# Indexation de vidéos et de maillages 3D dans le contexte MPEG-7

Titus Zaharia

**UNIVERSITE RENE DESCARTES - PARIS V**
**Centre Universitaire des Saints-Pères**
UFR DE MATHEMATIQUES ET INFORMATIQUE

***Thèse présentée en vue de l'obtention du grade de Docteur***
***de l'Université RENE DESCARTES - Paris V***

Discipline : Sciences de la Vie et de la Matière
Spécialité : Mathématiques – Informatique

Par

Titus Bogdan ZAHARIA

Sujet de la Thèse :

# Indexation de vidéos et de maillages 3D dans le contexte MPEG-7

Soutenue le 18 décembre 2001,

devant le jury composé de :

| | | |
|---|---|---|
| Monsieur le Professeur | Georges STAMON | Président |
| Monsieur le Professeur | Fernando PEREIRA | Rapporteur |
| Monsieur le Professeur | Philippe SALEMBIER | Rapporteur |
| Monsieur le Professeur | Frédéric TRUCHETET | Rapporteur |
| Madame le Professeur | Françoise PRÊTEUX | Directeur de Thèse |
| Monsieur le Professeur | Vasile BUZULOIU | Examinateur |
| Madame le Professeur | Christine GRAFFIGNE | Examinateur |

# Remerciements

Débuté il y a un peu plus de trois ans, ce travail n'aurait pu être conduit à son terme sans l'aide et la bonne volonté de nombreuses personnes.

C'est tout d'abord à Madame le Professeur Françoise Prêteux, responsable de l'Unité de Projets ARTEMIS de l'INT et directeur de cette thèse, que je tiens à exprimer mes respectueux remerciements. Qu'elle soit assurée de ma reconnaissance pour la qualité de la formation dont elle m'a fait bénéficier, pour les conseils et l'aide efficace qu'elle m'a généreusement offerts tout au long de ces années, pour son sens de la Solidarité, pour la confiance qu'elle m'a accordée dont je suis très touché et pour son exemplaire façon de me stimuler.

A Monsieur Georges Stamon, Professeur à l'Université Paris V, qui m'a fait l'honneur de présider ce jury, je tiens à témoigner mon respect et mes chaleureux remerciements.

A Monsieur Fernando Pereira, Professeur à *l'Instituto Superior Técnico*, qui retrouvera dans cette thèse une bonne partie de la terminologie MPEG-7 dont il est responsable, je tiens à exprimer mon respect et ma sympathie. Qu'il soit aussi assuré de toute ma reconnaissance pour l'intérêt qu'il a bien voulu porter à ce travail.

A Monsieur Philippe Salembier, Professeur à *l'Universitat Politecnica de Catalunya*, qui m'a fait l'honneur de s'intéresser à ce travail et d'y apporter le jugement du *Chair* du groupe MPEG-7 *Multimedia Description Schemes*, je tiens à exprimer ici mes chaleureux remerciements et mon entière gratitude.

A Monsieur Frédéric Truchetet, Professeur à l'Université de Bourgogne, qui a accepté la lourde tâche d'être rapporteur, je souhaite exprimer ici toute ma reconnaissance et tout le plaisir que j'ai pris dans les nombreux échanges que nous avons eus.

A Madame Christine Graffigne, Professeur à l'Université Paris V, je souhaite exprimer toute mon amicale gratitude pour l'aide indéfectible apportée pendant toutes ces années.

A Monsieur Vasile Buzuloiu, Professeur à l'Université POLITEHNICA de Bucarest, qui m'a fait l'honneur de participer à ce jury, je souhaite témoigner mon respect et toute ma reconnaissance. Accueilli avec amitié dans son équipe de traitement d'images dès le début de ma carrière en 1995, j'ai eu le privilège de bénéficier de sa formation et de ses conseils aussi bien sur le plan scientifique que sur le plan humain. Qu'il soit aussi assuré ici de mon admiration pour l'énorme travail qu'il conduit inlassablement depuis des années dans plusieurs universités roumaines et pour sa confiance dans la jeunesse, qu'il ne cesse de promouvoir et soutenir avec enthousiasme et obstination.

Je ne saurai oublier qu'il y a quatre ans déjà, j'ai eu le bonheur d'apprendre que mon dossier de candidature à une bourse de thèse en cotutelle avait été retenu par les experts du Service Culturel Français en Roumanie. En me soutenant dans ma formation doctorale, ils sont certainement à l'origine de ces travaux. Que Son Excellence, l'Ambassadeur de France en Roumanie, soit ici assuré de ma profonde et respectueuse reconnaissance et que Madame l'Attachée de Coopération Scientifique et Technique du Service de Coopération et d'Action Culturelle de l'Ambassade de France en Roumanie, accepte l'expression de mes vifs remerciements pour l'accueil qu'elle m'a réservé.

Je tiens aussi à associer à mes remerciements :

- Monsieur Jean-Marie Becker, Maître de Conférences à CPE Lyon, qui a guidé mes premiers pas de chercheur lors de mon projet de fin d'études, et m'a fait découvrir la très géométrique beauté de ses approches,
- Mesdames Elena Dobrescu et Elena Răileanu, Professeurs de mathématiques, pour la remarquable qualité didactique de leurs enseignements, dispensés avec autant de dévouement que d'abnégation aux jeunes bucarestois,
- Monsieur Nicolas Rougon, Maître de Conférences à l'INT, pour les moments privilégiés passés à discuter amicalement,
- Mesdames Nicole Teste et Evelyne Taroni, pour leur application constante à minimiser les petits soucis administratifs, et pour leur soutien quotidien plein de bonne humeur et d'humour,
- Madame Jeannine Clément, pour toute sa gentillesse et sa confiance.

Enfin, je souhaiterais remercier mes amis, qui de près ou de loin, par leur aide, esprit, humour, sourire, finesse... m'ont supporté tout au long de ces années : Răzvan Beuran, Mihai Ciuc, Mircea Curilă, Sorin Curilă, Ionuț Deaconeasa, Florin Dumitru, Cătălin Fetița, Mihai Ivanovici, Marius Malciu, Carlos Martin, Mihai Mitrea, Gérard Mozelle, Tudor Murgan, Caroline Petitjean, Elisabeta Podaru, Marius Preda, Dan Simion, Constantin Vertan.

*A mes parents*

# Table des matières

# Liste des figures

# Liste des tableaux

# Introduction

*"Bienvenue, ô voyageur, dans ce modeste hexagone. Si mes délires te semblent obscurs, ne t'imagine point pour autant que je raconte n'importe quoi, car je ne puis combiner une série quelconque de caractères, par exemple*

*dhcmrlchtdj*

*que la divine Bibliothèque n'ait déjà prévue, et qui dans quelqu'une de ses langues secrètes ne renferme une signification terrible."*

Jorge Louis Borges, "La Bibliothèque de Babel"[1]

Avec l'apparition de l'Internet, vers la fin des années quatre vingts, une énorme masse de données a littéralement envahi notre quotidien. Jour après jour, ce volume d'information ne cesse de croître, relayé depuis des millions de sites répartis dans le monde entier. Devant cette explosion informationnelle, les sentiments sont partagés entre angoisse des labyrinthes borgesiens et enthousiasme de découvrir une nouvelle dimension de la liberté, bien au-delà des frontières physiques, géopolitiques et linguistiques. Que nous l'aimons ou non, l'Internet *est.* Son existence est en train de révolutionner notre vision du monde et nous lance chaque jour de nouveaux défis.

Le volume croissant de données numériques aussi diverses qu'images fixes, vidéos, données 3D... requiert en effet de disposer de méthodes efficaces d'accès, de manipulation et de structuration de ces contenus multimédias hétérogènes et complexes. La solution classique, héritée de la gestion des bases de documents de type texte, renvoie à des méthodes d'annotation textuelle, et consiste à associer aux données des mots clé, décrivant d'une façon plus ou moins synthétique le contenu. Au-delà des efforts de nombreuses équipes de documentalistes pour créer manuellement les annotations, c'est le caractère fortement subjectif de celles-ci qui porte les principales limitations d'ordre sémantique et linguistique de cette méthodologie. Les *Internautes* d'aujourd'hui sont d'ailleurs constamment confrontés à ce problème, largement illustré par Umberto Eco à travers les péripéties d'un érudit de notre temps en promenade sur Internet[2].

Les paradigmes de la représentation et de l'accès par le contenu des données multimédia semblent alors offrir le bon cadre pour indexer d'une façon objective et homogène des contenus aussi riches que proteïformes. Il s'agit ici d'associer aux données multimédias non plus des annotations textuelles, mais des signatures spécifiques, liées aux caractéristiques visuelles intuitives d'image comme la couleur, la texture, la forme et le mouvement.
Cette approche renvoie au vaste domaine de l'analyse d'image et nécessite :

- d'identifier de façon systématique et objective les éléments signifiants d'un document multimédia,
- de déterminer des signatures pertinentes et de leur associer des mesures de similarité spécifiques afin de les comparer efficacement pour effectuer des requêtes,

[1] Jorge Louis Borges, "Fictions", Gallimard, Paris, 1951.
[2] Umberto Eco, "Comment voyager avec un saumon", Grasset et Fasquelle, Paris, 1998.

- d'exprimer, de stocker et de transmettre les différentes descriptions dans un format approprié et universel,
- d'élaborer les outils afférents de visualisation, navigation, annotation par le contenu et requêtes par similarité permettant à l'utilisateur un accès aisé aux contenus indexés.

Par leur caractère objectivement lié aux données, les représentations par le contenu offrent de bonnes prémices d'interopérabilité des descriptions. Identifiant les grands enjeux socio-économiques pour disposer de technologies standardisées de description de données multimédias, le groupe MPEG (*Moving Pictures Expert Group*) de l'ISO (*International Standardization Organization*) a lancé en février 1999 un appel à propositions, pour élaborer le futur standard de description des contenus "*Multimedia Content Description Interface*", brièvement appelé MPEG-7. Le très vif intérêt rencontré parmi les principaux acteurs, qu'ils soient industriels ou académiques, du monde du multimédia, s'est concrétisé par plus de 650 propositions soumises en réponse à cet appel. Cette thèse s'inscrit pleinement dans le cadre du processus de standardisation MPEG-7, auquel nous avons participé activement depuis son début.

Le premier chapitre du mémoire précise tout d'abord le contexte international de la normalisation dans lequel s'inscrivent les recherches et développements que nous avons réalisés dans le domaine de l'indexation des objets visuels, en en précisant notamment règles, procédures et calendrier. Ensuite, une analyse synthétique de la première version du futur standard MPEG-7 sur la description des contenus multimédias est présentée, en termes de descripteurs, schémas de description, langage de description de données, et applications et services supportés. Enfin, les descripteurs visuels de couleur, texture, forme et mouvement retenus sont détaillés, en mettant l'accent sur leur principe et leur interprétation mathématique.

Le deuxième chapitre traite de l'indexation des documents vidéos dans le cadre de la modélisation paramétrique de mouvement, en commençant par présenter d'une manière comparée les différents modèles classiquement utilisés et les algorithmes d'extraction, robustes ou non-robustes, associés. Malgré la compacité et la fidélité de ces représentations, leur exploitation directe dans le cadre d'applications de requêtes par similarité est restée peu explorée, par manque de mesures de similarité appropriées et de mécanismes de gestion du caractère dynamique des données vidéos. Afin de lever ce verrou technologique, nous proposons une famille de mesures de similarité, définie dans l'espace des champs de vitesse (MSCV). Elle dépend d'une fonction distance générique spécifiée en fonction du type de requête. Les problèmes d'optimisation en temps de calcul, d'alignement spatio-temporel et de pondération des composantes translationnelle et homogène de mouvement sont analysés et une solution mathématique proposée et mise en œuvre sous forme d'une famille de mesures de similarité simplifiées exprimées dans l'espace des champs de vitesse (MSSCV).
Les expérimentations conduites sur des bases de test synthétique et naturelle avec vérité terrain démontrent, objectivement et quantitativement, par estimation du critère *Bull Eye* retenu dans le cadre de MPEG-7, la nette supériorité des MSSCV par rapport aux mesures de similarité définies dans l'espace des paramètres de mouvement et établit ainsi la pertinence du descripteur de mouvement paramétrique que nous avons proposé et qui est retenu dans MPEG-7.

Le troisième chapitre concerne l'indexation d'objets 3D maillés à l'aide de descripteurs de forme (DF), sous contraintes d'invariance géométrique et de robustesse topologique.
Le spectre de forme 3D (SF3D), que nous avons proposé et qui est retenu comme DF dans MPEG-7, est tout d'abord introduit. Le SF3D est défini comme la distribution d'un index de forme caractérisant localement la géométrie d'une surface 3D et est exprimé comme la coordonnée angulaire de la représentation polaire du vecteur de courbures principales. Intrinsèquement invariant aux transformations géométriques, le SF3D n'est pas robuste vis-à-vis des multiples représentations topologiques d'un même objet.
C'est pourquoi un nouveau DF, intrinsèquement stable topologiquement, est proposé. Dérivé de la transformée de Hough 3D, le descripteur de Hough 3D (DH3D) n'est en revanche pas invariant aux transformées géométriques. Nous montrons mathématiquement comment il peut être associé de façon optimale en termes de compacité de représentation et de complexité de calcul à une procédure d'alignement spatial lui conférant alors un comportement d'invariance géométrique. Cela conduit à définir le DH3D optimal (DH3DO).
Après avoir spécifié les mesures de similarité utilisées lors des applications de requête et avoir décrit la base de 1300 modèles utilisée, les deux DF sont évalués et comparés objectivement en terme de score *Bull-Eye*, ce critère établissant une nette supériorité du DH3DO.

Enfin, le dernier chapitre présente le système d'indexation multimédia AMIS (*Advanced Multimedia Indexing System*), qui offre dans le cadre du standard MPEG-7, une plate-forme de visualisation/navigation/annotation de contenus audio-visuels naturels et synthétiques. Ce système, fondé sur une architecture modulaire extensible, intègre les divers types de médias (images fixes, séquences d'images, documents vidéos, maillages 3D) sous différents formats et offre un ensemble diversifié d'outils de segmentation temporelle et spatio-temporelle et d'extraction des descripteurs MPEG-7 pour des applications de recherche par le contenu. En particulier, le principe d'interopérabilité devient effectif par la mise en œuvre des schémas de description MPEG-7 qui permettent de s'affranchir des problèmes de cohérence, d'héritage, de hiérarchie, de synchronisation et d'intégration de signatures hétérogènes.
L'ensemble logiciel ainsi réalisé démontre pour la première fois en grandeur réelle, pour des applications d'indexation multimédia, comme l'archivage vidéo, la vidéo cliquable et la reconnaissance des gestes en langue des signes française le caractère effectivement opérationnel de schémas de description audio-visuels génériques, normalisés MPEG-7.

Enfin, ce mémoire se conclut en synthétisant nos contributions originales et en esquissant quelques perspectives tant méthodologiques qu'en termes d'applications et de nouveaux services multimédias.

# Chapitre 1

# Normalisation de l'image : de la compression vidéo aux représentations par le contenu


## Résumé

*Ce chapitre a pour objectifs :*

- *De préciser le contexte international de la normalisation dans lequel s'inscrivent les recherches et développements que nous avons réalisés dans le domaine de l'indexation des objets visuels, en en précisant notamment les règles, procédures et le calendrier,*
- *De présenter une analyse synthétique de la première version du futur standard MPEG-7 sur la description des contenus multimédias en termes de descripteurs, schémas de description, langage de description de données, et applications et services supportés,*
- *De détailler les descripteurs visuels de couleur, texture, forme et mouvement retenus en mettant l'accent sur leur principe et leur interprétation mathématique.*



**Mots Clef**

*International Standardization Organization, Moving Picture Expert Group, indexation par le contenu, représentations orientées objet, métadonnées, descripteur, schéma de description, langage de description de données, attributs d'image.*

## 1.1 Introduction à la normalisation de l'image

Le terme *normalisation de l'image*, consacré par l'AFNOR (Association Française de NORmalisation), est une formulation compacte qui veut dire que l'on s'attaque à la représentation des données de type image.

Le schéma synoptique de l'organisation de la normalisation de l'image, présenté Figure 1.1, se décline à différents niveaux internationaux d'instances représentatives. Les sigles utilisés, de façon standard, sont les suivants :

- ISO : International Standardization Organization,
- IEC : International Electrotechnical Commission,
- TC : Technical Committee,
- JTC1 : Joint Technical Committee N°1,
- SC : Sub-Committee,
- WG : Working Group.

Précisons que les SC sont structurés en un certain nombre de WG. En ce qui concerne le SC29, il regroupe les activités de 3 WG :

1. WG1 : JPEG (Joint Picture Expert Group),
2. WG11 : MPEG ( Moving Picture Expert Group),
3. WG12 : MHEG (Multimedia and Hypermedia information coding Expert Group - dissout à Singapour, Mars 2001).

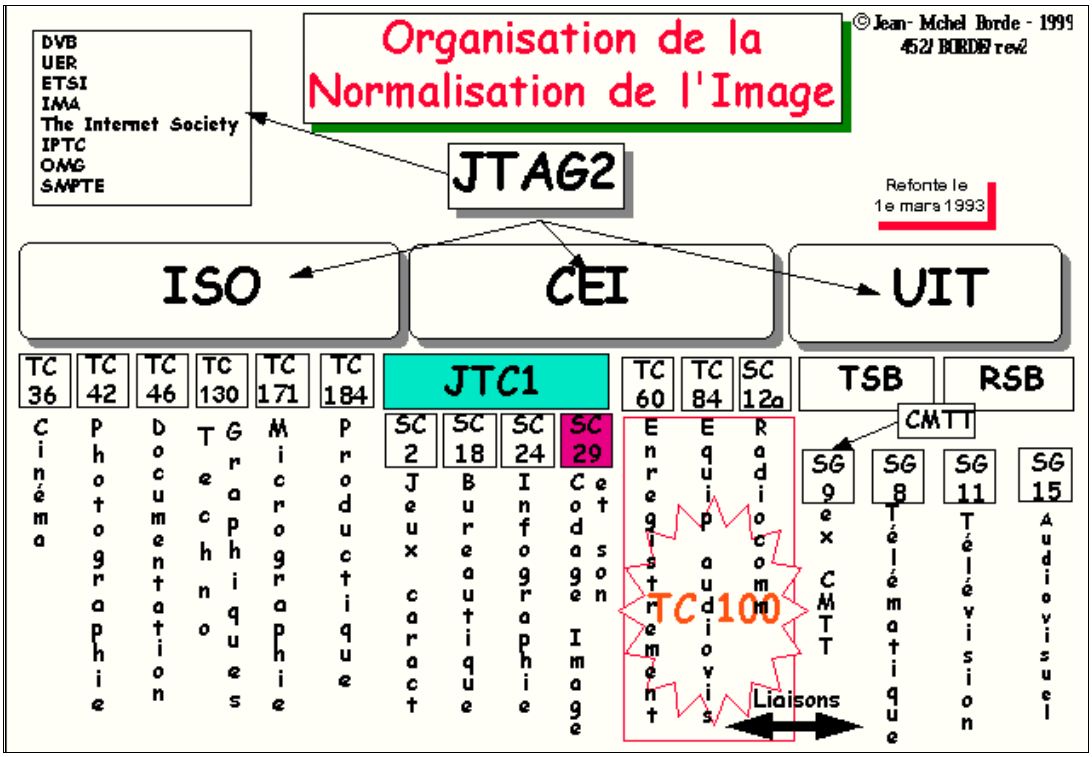



**Figure 1.1.** Schéma synoptique de l'organisation de la normalisation de l'image.

L'ensemble de ces instances a pour objectif de produire des normes, *i.e.* des documents papier comportant :

- une référence,
- une date d'édition,
- un titre,
- une marque d'appartenance.

Ces mêmes caractéristiques se retrouvent sur les CD actuellement associés à une norme (Figure 1.2).

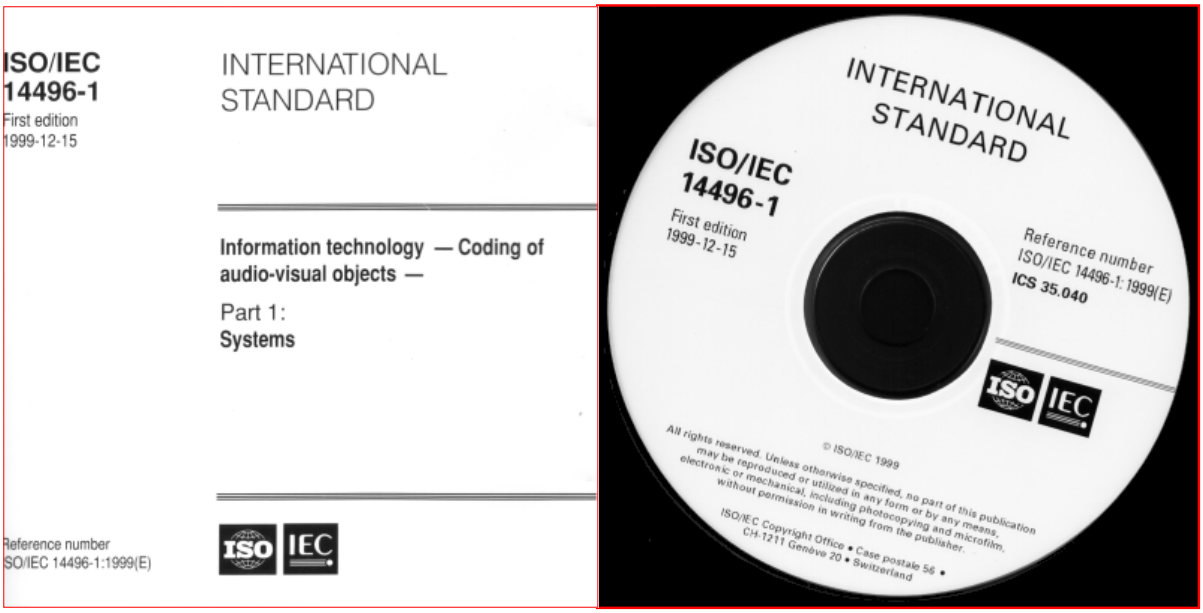



**Figure 1.2.** Exemple de référence d'une norme.

Une norme, plate-forme d'interopérabilité au service d'applications, est le résultat d'un travail collaboratif impliquant industriels, représentants académiques, consommateurs ou prescripteurs et les pouvoirs publics. Elle représente le consensus des personnes impliquées sur un ensemble de technologies. Une norme sert à la mise en fabrication de produits industriels et à la vérification de leur conformité. C'est également, en un certain sens, un bon modèle de l'état de l'art dans le domaine concerné. Toutefois, des procédures de maintenance sont prévues de manière à faire évoluer le standard en lui permettant d'intégrer de nouvelles technologies plus performantes, faisant ainsi bénéficier la norme des progrès scientifiques réalisés dans des communautés aussi bien industrielles qu'académiques. C'est ainsi que MPEG-4 vient d'avoir la possibilité, suite à différents *Call for Proposals* (CfP) [MPEG-4-VideoCfP, MPEG-4-VideoCfP] d'intégrer des technologies plus performantes de codage audio et vidéo et d'étendre les technologies SNHC aux objets 3D déformables. Dans le même esprit de progrès technologique, MPEG-4 vient récemment de décider une nouvelle activité de compression [MPEG-Press07.01] conjointement avec le comité ITU-T SG16 (*International Telecommunication Union - Telecommunication Standardization Sector, Study Group 16*), qui se concrétisera dans une nouvelle version du standard MPEG-4, dont la sortie est prévue au printemps 2003.

Les différentes étapes du processus de normalisation s'organisent sous forme d'un triptyque : confrontation (stade actuel de MPEG-7 version 2), convergence (stade actuel de MPEG-4) et évolution/maintenance (stade actuel de MPEG-1 et MPEG-2). Ces trois volets se déclinent au sein de réunions internationales (environ 4 à 5 par an) au niveau des WG.

Comment participer à ces réunions ? La procédure à suivre est simple et est précisée par la résolution 30 du SC29 du 22-24 Mars 1999 rappelée dans l'Annexe 1.1.

Comment se déroulent les réunions ? Suivant des principes quasi-immuables au sein du WG11. Avant la réunion, les contributions des différents acteurs sont enregistrées électroniquement (procédure technique spécifique par numéro d'enregistrement ISO). Pendant la réunion (du lundi au vendredi), elles sont examinées, discutées, soit au sein de petits groupes spécifiques (*Ad Hoc Group*), soit au sein des groupes pléniers qui structurent les activités MPEG (ceux-ci peuvent différer selon qu'il s'agit de MPEG-4 ou de MPEG-7). Une décision consensuelle est alors adoptée par le groupe concerné (Figure 1.3) : rejet, demande de modification, recommandation d'association avec des propositions similaires ou complémentaires, intégration au sein d'un CE[1] (*Core Experiment*) pour évaluation technique ou création d'un CE spécifique, intégration dans le XM[2] (*eXperimentation Model*) ou le VM[3] (*Validation Model*).

En fin de semaine, des documents de sortie, référencés ISO, dressent le bilan cumulé des activités en termes de résolutions, décisions, recommandations, mandatements, spécifications, évaluation de technologies… Certains de ces documents sont publics, d'autres restent internes au WG11.

[1] CE (terminologie héritée des divers standards MPEG) : l'objectif est de choisir entre des propositions concurrentes, l'expérimentation étant conduite par au moins deux parties indépendantes dans des conditions pré-définies et pour des critères d'évaluation fixés.

[2] XM (terminologie spécifique à MPEG-7) : environnement contrôlé pour y développer les expérimentations contenant *plus* que les technologies du standard (moteur de recherche, extraction de primitives…).

[3] VM (terminologie spécifique à MPEG-4) : environnement contrôlé pour y développer les expérimentations des technologies du standard.

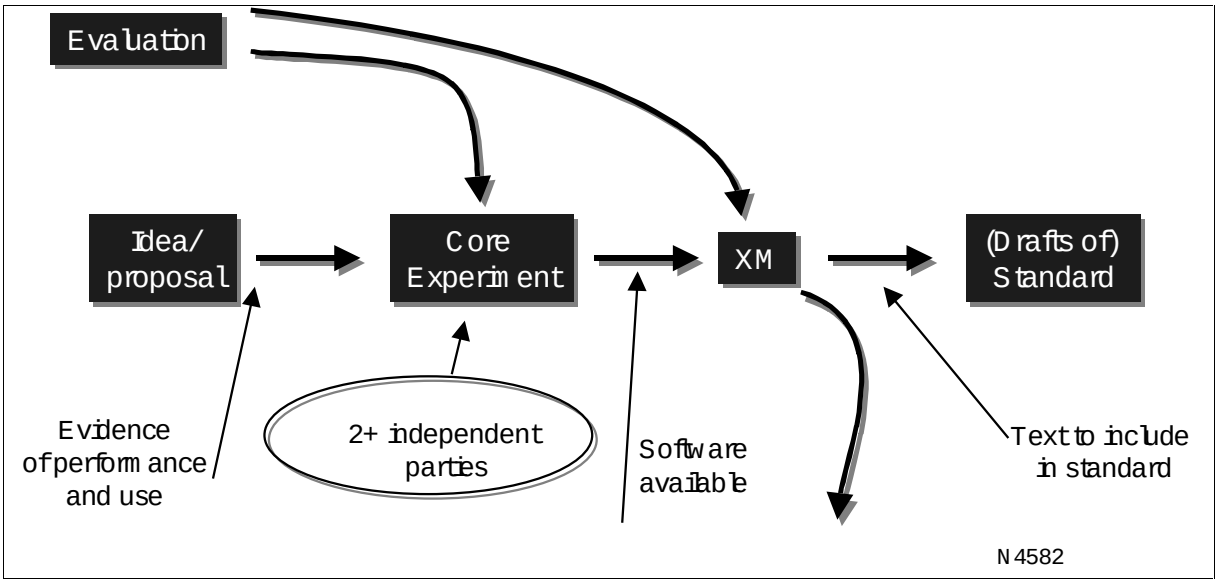



**Figure 1.3.** Des propositions au standard.

Ces réunions contribuent donc à l'élaboration des futurs standards suivant un calendrier officiellement déposé auprès du SC29 qui le soumet pour approbation aux échelons supérieurs. La progression vers le standard s'effectue par étapes (Figure 1.4) :

- *Working Draft* (WD) correspondant aux éditions successives du «brouillon» de la norme,
- *Committee Draft* (CD) donnant une première version stabilisée dans ses grandes orientations de la norme mais encore incomplète techniquement,
- *Final Committee Draft* (FCD) intégrant via les commentaires des représentants nationaux (*National Bodies* - NB) des modifications de technologies non fondamentales par rapport au CD et des corrections techniques,
- *Final Draft International Standard* (FDIS) préfigurant, sous forme d'une version encore amendable (à l'extrême marge) via les commentaires des NB, le futur standard,
- *International Standard* (IS) spécifiant le standard.

A partir du CD, les votes s'effectuent via les représentants nationaux, *i.e.* l'AFNOR pour la France. Ces votes correspondent aux recommandations des experts du pays.

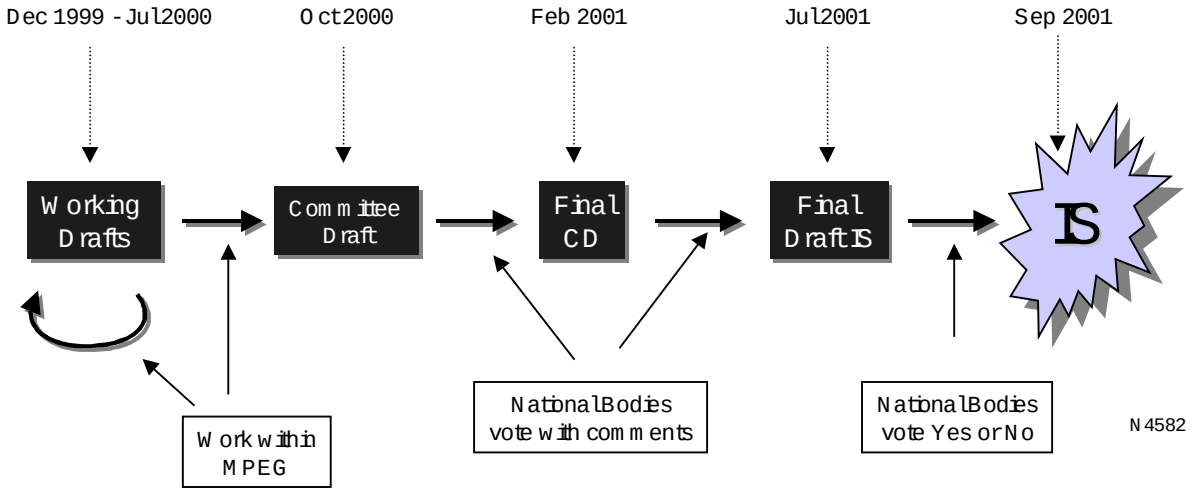



**Figure 1.4.** Les différentes étapes vers le standard et les acteurs impliqués (calendrier de MPEG-7).

Après cette brève introduction au contexte de la normalisation, ce chapitre présente une synthèse des activités du WG11, notamment au travers de ses développements dans le cadre de MPEG-4 et de MPEG-7. Pour toute information sur les normes MPEG-1 et MPEG-2, le lecteur est invité à se reporter aux documents ISO 11172-1/5 et 13818-1/9, respectivement.

Le standard MPEG-4 marque en effet un saut décisif dans l'évolution des technologies de codage vidéo, en adoptant notamment les paradigmes d'une seconde génération d'outils de compression qui s'appuient sur des représentations orientées objet (naturel et de forme arbitraire ou synthétique), des principes de codage sélectif, la compression des paramètres d'animation des avatars, et le support d'une gamme diversifiée de types de média et le multimédia interactif. Détailler les technologies de codage MPEG-4 sort du contexte de ce travail, orienté vers les représentations par le contenu d'objets multimédias. Toutefois, MPEG-4, par le large éventail de fonctionnalités supportées révolutionne complètement le monde du multimédia numérique [Koenen97, Pereira00]. Dépassant largement le contexte *stricto sensu* d'un standard de compression vidéo, il offre les prémices des représentations par le contenu (*Object Content Information*) au sein du groupe MPEG. C'est pour cette raison que nous proposons ci-dessous un aperçu très synthétique du standard MPEG-4.

## 1.2 Le standard MPEG-4

MPEG-4 est un standard ISO/IEC (depuis décembre 1999 en version 1) développé par le WG11, qui a déjà produit MPEG-1 et MPEG-2, standards qui ont rendu possibles la vidéo interactive sur cédéroms et la télévision numérique. MPEG-4 traite des objets audiovisuels 2D/3D naturels et/ou synthétiques et décline des objectifs de codage sélectif et de composition de scènes intégrés dans un système offrant des outils génériques et des fonctionnalités nouvelles d'accès universel et d'interactivité.

## 1.2.1 MPEG-4 : principes et applications

MPEG-4 est dédié aux champs d'application suivants :

- télévision numérique,
- graphiques interactifs (contenus synthétiques),
- multimédia interactif (Internet, Intranet).

MPEG-4 contient toutes les technologies qui permettent l'intégration de la production, de la distribution et du paradigme d'accès aux contenus de ces trois domaines.

Le paradigme d'interaction et accès au contenu s'exprime à l'aide de nouveaux outils permettant la manipulation d'objets vidéos de forme arbitraire, l'édition de flux binaires et la scalabilité par objet vidéo. Chaque objet vidéo est codé sur plusieurs niveaux (*layers*) : un niveau de base (*base layer*), suivi de plusieurs *enhancement layers*, correspondant à des degrés de qualité et complexité croissants. MPEG-4 intègre dans un schéma de codage générique, trois types différents de scalabilité : temporelle (contrôle du *frame rate*), spatiale (contrôle de la résolution spatiale) et en qualité (rapport signal / bruit). Ainsi, un seul flux binaire MPEG-4 peut-il être réutilisé sous diverses contraintes de largeur de bande et de complexité. Enfin, MPEG-4 offre les outils spécifiques au paradigme d'accès universel, en conjuguant le principe de scalabilité par objet vidéo à des techniques permettant la transmission sur des canaux de communication bruités (*cf.* réseaux mobiles), et en offrant des outils tels que l'insertion de marqueurs de re-synchronisation, partitionnement des données, codes réversibles ou rafraîchissement intra.
Grâce à MPEG-4, les auteurs peuvent créer des contenus réutilisables de façon plus souple qu'avec les formats non intégrés existants. MPEG-4 permet également une gestion des droits d'accès et une protection contre la copie.
MPEG-4 offre aux fournisseurs de services réseaux et aux distributeurs un cadre garantissant une *quasi* décorrélation des contenus vis-à-vis des technologies de distribution (réseaux et diffusion). Par exemple, MPEG-4 rend possible l'optimisation de la qualité de service tout au long de la distribution sur un réseau hétérogène.
MPEG-4 propose aux utilisateurs l'intégration de nombreuses techniques avancées : consultation de contenus sur des terminaux variés du téléphone portable amélioré au PC générique en passant par la télévision avec un décodeur, accès interactif à des applications (par opposition à la consultation d'émissions linéaires)…

MPEG-4 réalise ces performances en fournissant des méthodes standardisées pour :

1. **représenter** de manière compacte des "atomes" de contenus audios, vidéos ou audiovisuels appelés "objets". Ces objets peuvent être d'origine naturelle (un son ou une vidéo enregistrée) ou synthétique (voix synthétique, musique MIDI, scène 3D VRML) ;
2. **composer** ces objets afin de créer des objets audiovisuels composites appelés "scènes" ;
3. **multiplexer** et **synchroniser** les données associées aux objets, pour leur transport sur un réseau avec une qualité de service adaptée à la nature de chacun d'eux ;

4. **interagir** avec la scène reconstituée au niveau du récepteur.

## 1.2.2 Structuration de MPEG-4

Le groupe MPEG-4 est organisé en cinq groupes de travail.

Le groupe *Vidéo* est dédié à la compression et la représentation d'objets vidéo d'origine naturelle.

Le codage vidéo rassemble des outils aussi divers que ceux dédiés au codage de la forme, de la texture, des vecteurs de mouvement et des textures fixes. La Figure 1.5 présente le schéma synoptique pour le codage vidéo.

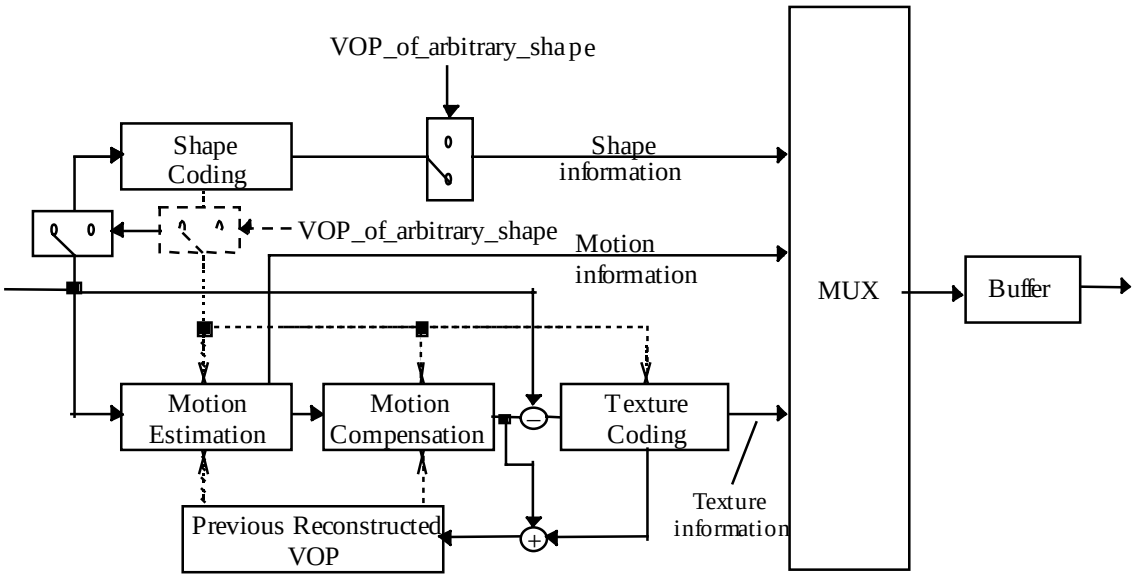



**Figure 1.5.** Schéma de codage MPEG-4 [MPEG-4].

Notons qu'un des éléments nouveaux de ce schéma est le codage prédictif de la forme, réalisé aussi bien en mode intra (sans compensation de mouvement) qu'en mode prédictif. Il repose sur un codeur arithmétique avec contextes. L'estimation du mouvement, réalisée avec une précision allant jusqu'au quart de pixel, est fondée sur le principe de *block matching* qui consiste à minimiser le critère défini par la somme des valeurs absolues des erreurs de compensation. Les vecteurs de mouvement, qui peuvent être associés aux blocs (8x8 pels) ou aux macro-blocs (16x16 pels) sont codés par la suite de manière différentielle. Le codage de la texture est fondé sur la transformation Cosinus discrète (*Discrete Cosine Transform* – DCT) et s'applique soit aux erreurs résiduelles de prédiction après la compensation du mouvement, dans le cas du codage prédictif, soit directement à la texture initiale, dans le cas du mode intra. Des tables de quantification permettent la quantification linéaire ou non-linéaire des coefficients DCT suivant les fréquences spatiales correspondantes.

Les textures fixes réclament un degré plus avancé de scalabilité et font appel aux transformations en ondelettes. Pour des raisons d'efficacité en termes de performance de compression, les ondelettes biorthogonales de Daubechies (9-3) ont été choisies. Les coefficients obtenus par la décomposition hiérarchique pyramidale du signal sont par la suite codés en utilisant l'algorithme *de Zero Tree Wavelet Coding* (ZTW).

Des performances élevées sont obtenues en utilisant le concept de mosaïques (*sprites*). Une mosaïque est codée et transmise une seule fois, les trames individuelles étant recréées au niveau du récepteur par découpage de la mosaïque en fonction des paramètres de transformation.

Les autres groupes impliqués dans MPEG-4 sont les suivants :

1. *Audio* concerne le domaine audio et inclut compression, représentation et composition, y compris pour certains aspects synthétiques.
2. *Synthétique (SNHC)* développe des méthodes de compression et de représentation pour des objets visuels synthétiques ou hybrides, statiques ou animés (notamment visages ou corps humains), ainsi que pour la synthèse de parole.
3. *DMIF* (*Delivery Multimedia Integration Framework*) élabore une couche d'abstraction rendant l'application MPEG-4 indépendante du réseau ou de la méthode de distribution.
4. *Système* intègre à la fois les définitions :
    (i) de la composition spatiale, temporelle et interactive des objets définis par les autres groupes, afin de construire une scène,
    (ii) du catalogue des objets utilisés et de la signalisation de leurs caractéristiques,
    (iii) d'un multiplex (optionnel),
    (iv) d'un modèle de récepteur permettant de limiter la quantité de mémoire tampon nécessaire au bon fonctionnement d'un récepteur.

La définition d'une scène MPEG-4 est effectuée à l'aide du format BIFS (*BInary Format for Scenes*), version binaire d'un format fortement inspiré de VRML'97. Les fonctionnalités de BIFS, issues de VRML ou spécifiques à MPEG, sont :

- la représentation et la composition d'objets 3D (VRML),
- la représentation et la composition d'objets 2D,
- la représentation de scènes animées pour lesquelles l'information d'animation est contenue dans la scène au départ (VRML),
- la représentation de scènes dynamiques et animées : les modifications de la scène peuvent être transmises au cours du temps, voire générées en fonction d'actions de l'utilisateur,
- la transmission progressive de scène,
- la définition d'interactivité à l'aide de la souris, ou plus généralement d'un pointeur (VRML),
- la définition de scripts (langage ECMAScript) pour étendre les fonctions d'interactivité.

L'interactivité dans une scène BIFS résulte de l'utilisation de *capteurs (sensors)*, d'*interpolateurs* et de *routes*. Un capteur est un objet (par exemple *TouchSensor*) qui produit des événements en fonction de l'action de l'utilisateur sur un élément d'interface (comme la souris). Ces événements sont transmis à d'autres objets par l'intermédiaire de routes. L'événement généré peut être utilisable directement par le destinataire de l'événement, ou peut avoir besoin d'être adapté. Un premier niveau d'adaptation suppose le transit de l'événement par un interpolateur, qui par exemple traduit un événement à valeur flottante en un événement à valeur vectorielle permettant de déplacer un objet. Un second niveau d'adaptation, plus complexe, suppose l'utilisation de scripts.

Un aspect très important de MPEG-4 concerne la performance en terme de compression de données audiovisuelles. Un des succès les plus importants de la norme MPEG-2 est de stocker un film sur un support DVD, avec une qualité visuelle meilleure que celle d'une cassette vidéo. Les performances en compression de MPEG-4 ont encore été améliorées : un film de 1h30 est stocké sur un seul CD avec une qualité supérieure à celle de MPEG-2. En outre, les performances de MPEG-4 ont fait l'objet d'extensions importantes au niveau de :

- leurs champs d'application, en particulier avec la conservation d'une bonne qualité à des débits beaucoup plus faibles, par exemple pour la visiophonie ou la transmission de voix à très bas débit ;
- nouveaux types d'objets à coder, comme les objets synthétiques ;
- l'introduction de fonctionnalités nouvelles : robustesse aux erreurs de transmission, codage de forme non rectangulaire pour les vidéos, scalabilité.

Mentionnons enfin qu'au cours de la spécification de ce programme pour le moins ambitieux, le comité MPEG a été amené à définir des priorités et pour chacune d'elles, des échéances différentes. Ces priorités se traduisent par trois versions du standard à paraître à un an d'intervalle, dont les contenus sont synthétisés dans l'Annexe 1.2.

La richesse de l'approche MPEG-4 est d'autant plus manifeste qu'elle contient intimement en elle les germes des évolutions futures des technologies standardisées MPEG.
En effet, un flux MPEG-4 est un contenu vidéo enrichi de divers éléments d'information relatifs aux différents objets individuels considérés, comme durée de vie, régions support... De plus, MPEG-4 définit des mécanismes d'accès manuel et d'interaction avec l'utilisateur à l'aide des interfaces munies de capteurs spécifiques.
Il vient tout naturellement à l'esprit la possibilité d'enrichir encore davantage cette représentation, en associant aux différents objets des descripteurs spécifiques débouchant sur des fonctionnalités nouvelles, comme par exemple l'accès automatique et les requêtes par le contenu.

Ces différents aspects ont donc été développés *in extenso* par le groupe MPEG, dans le cadre générique et par ailleurs complètement indépendant de MPEG-4, de l'élaboration du futur standard "*Multimedia Content Description Interface*" (MCDI), brièvement appelé MPEG-7.

# 1.3 Panorama sur MPEG-7

Détaillons à présent les différents éléments de MPEG-7 que nous considérons dans ce travail, en adoptant un approche descendante. Commençons donc par énoncer les objectifs et les applications ciblés par MPEG-7.

## 1.3.1 Contexte et objectifs de MPEG-7

L'accroissement du volume des données numériques aujourd'hui accessibles sur l'Internet, via des bases de données ou la diffusion par les bouquets numériques, requiert de disposer de modalités d'accès intelligent à ces contenus multimédias qui sont composés d'images fixes, de vidéo, d'audio et de texte.
Cette nécessité fait écho à des enjeux socio-économiques importants, s'affirmant dans des contextes d'applications professionnels ou grand public aussi divers que les télécommunications (codage, téléports, réseaux...), les services en ligne (commerce électronique, informations personnalisées, ...) et la production audiovisuelle (télévision, industrie cinématographique, vidéo, post-production, archivage, accès public aux fonds collectifs...).

Le futur standard MPEG-7 a pour objectif [MPEG-7] de fournir des descriptions standardisées des contenus multimédias et de supporter un large éventail d'applications potentielles. MPEG-7 standardisera donc (Figure 1.6) :

1. **Un ensemble de descripteurs**
   Un Descripteur (*Descriptor - D*) est une représentation d'une primitive (*feature*) d'image. Un descripteur définit la syntaxe et la sémantique de la représentation de la primitive.

2. **Un ensemble de schémas de description**
   Un Schéma de Description (*Description Scheme - DS*) spécifie la structure et la sémantique des relations entre ses composantes, qui peuvent être aussi bien des Descripteurs que d'autres Schémas de Description.

3. **Un langage de définition de description**
   Ce langage (*Description Definition Language - DDL*), fondé sur *XML Schema* [XML], doit permettre de créer de nouveaux schémas de description, de nouveaux descripteurs et également d'étendre et de modifier des schémas de description existants.
   Mentionnons que la question d'adoption d'un langage de description a été précédemment considérée par MPEG, au niveau du standard MPEG-4. Il s'agit du format XMT (*Extensible MPEG-4 Textual Format*) [Kim00] qui définit un format textuel de représentation de scène, fondé sur XML et intégrant aussi bien des éléments de X3D (qui est la transcription en XML de VRML97, réalisée par le Consortium Web 3D) [X3D] que de SMIL (*Synchronized Multimedia Integration Language*) [SMIL], le langage de description que le Consortium W3C a développé spécifiquement pour la création des présentations multimédia synchronisées. XMT correspond en effet à une "xml-isation" des BIFS, assurant une relation biunivoque entre représentations binaires et textuelles. Par ailleurs, L'intégration du DDL MPEG-7 dans le cadre XMT a été également envisagée, pour des raisons d'interopérabilité augmentée. Dépassant largement le cadre de ce travail, cette problématique ne sera

pas détaillée plus avant dans ce mémoire.

### 4. Les schémas de codage (*Coding Schemes*)

Disposer de descripteurs et de schémas de description pose en pratique des problèmes de taille de représentations, de stockage et de transmission. Il est alors nécessaire de disposer de mécanismes adéquats de codage des divers descripteurs et schémas de description satisfaisant aux requêtes de type efficacité de la compression, résistance aux erreurs dans le cas des transmissions sur des canaux bruités, accès aléatoire, etc. MPEG-7 a actuellement retenu un mécanisme générique de codage, appelé BiM (*Binary format for Metadata*), qui associe de manière biunivoque à chaque description exprimée en langage de description MPEG-7 une représentation binaire compacte.

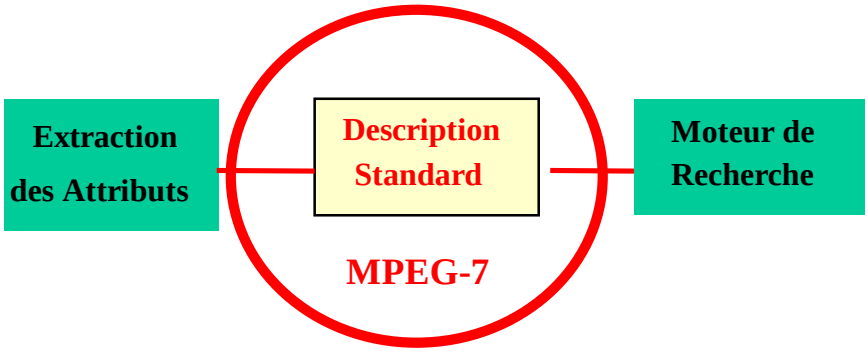



**Figure 1.6.** Le domaine de MPEG-7 (*Multimedia Content Description Interface*).

En l'état actuel des développements de MPEG-7, un schéma possible de la chaîne de traitement MPEG-7 est présenté Figure 1.7.

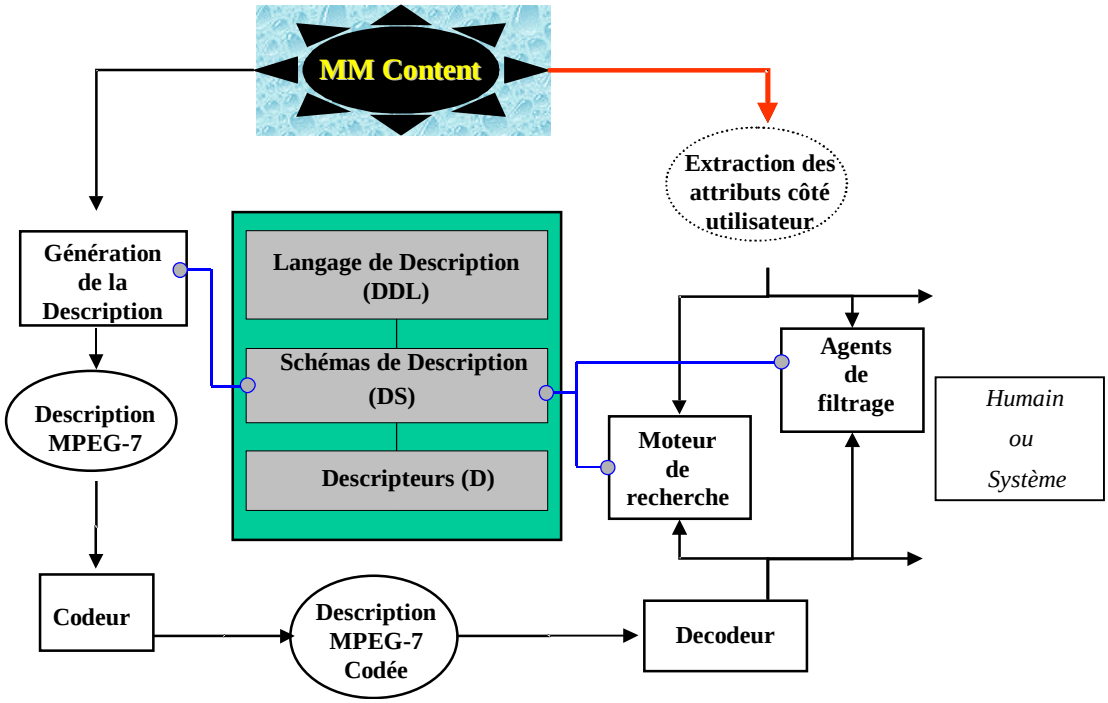



**Figure 1.7.** Schéma possible de la chaîne de traitement de MPEG-7 [MPEG-7REQ].

Soulignons que les méthodes d'extraction des descripteurs et les mesures de similarité associées restent en dehors du standard, qui se borne à quelques recommandations non normatives, incluses au niveau du logiciel de référence. Cette stratégie laisse la porte ouverte à de futures avancées méthodologiques dans ces domaines, sans pour autant remettre en question les technologies standardisées. Toutefois, dans le cas spécifique de MPEG-7, pour des raisons d'interopérabilité, une partie importante de l'extraction est définie par la sémantique de différentes composantes des descripteurs. Citons dès à présent comme exemples typiques, le descripteur de texture homogène (où les coefficients des filtres de Gabor, le nombre de sous-bandes spectrales et les mesures énergétiques associées sont figés, *cf.* Paragraphe 1.3.5.1) et le descripteur de reconnaissance de visage (où les 49 vecteurs propres sont fixés, *cf.* Paragraphe 1.3.5.6).

## 1.3.2 Domaines d'application de MPEG-7

MPEG-7 propose de couvrir une gamme d'applications aussi large que possible. Les divers domaines d'application ciblés par MPEG-7 et spécifiés dans le document "Applications" [MPEG-7Appl], renvoient :

- à l'archivage radio,
- à la TV et le cinéma,
- aux systèmes d'informations géographiques ou touristiques,
- au journalisme,
- à l'éducation,
- au monde des loisirs et divertissements (recherche de jeux, karaoke),
- aux services culturels (musée d'histoire, galeries d'art),
- à la télédétection (cartographie, écologie, gestion des ressources naturelles),
- à l'architecture,
- au domaine bio-médical,
- à la télésurveillance,
- au commerce électronique.

Les applications MPEG-7 sont structurées en deux classes : les applications de type *pull* et celles de type *push*.

La catégorie *pull* regroupe les applications qui ciblent l'extraction et la recherche de données audio-visuelles dans des bases de données, archives, ou sur *Internet*... Dans ce contexte, il est explicitement fait mention des applications suivantes :

- Stockage et requête de l'information dans des bases de données vidéos (*Storage and retrieval of video databases*),
- Fourniture d'images et de vidéos pour la production professionnelle de média (*Delivery of pictures and video for professional media production*),
- Applications commerciales musicales (*Commercial musical applications*),
- Bibliothèques d'effets sonores (*Sound effects libraries*),
- Bases de données de discours historiques (*Historical speech databases*),

- Recherche de films selon les événements audio (*Movie scene retrieval by memorable auditory events*),
- Enregistrement et recherche de marques enregistrées (*Registration and retrieval of mark databases*).

La catégorie *push* suit plutôt le paradigme du filtrage et de la sélection des données, pour des applications spécifiqes au monde du *broadcasting* ou du *webcasting* (encore en émergence). Pour illustrer l'éventail de ce type d'applications, le document MPEG fait référence :

- à la sélection et au filtrage des médias (*User agent driven media selection and filtering*),
- aux services TV personnalisés et présentations multimédias intelligentes (*Personalized Television Services, Intelligent multimedia presentation*),
- aux facilités d'accès à l'information pour les personnes handicapées (*Information access facilities for people with special needs*).

Enfin, MPEG-7 se propose également d'aborder un certain nombre d'applications ayant un profil hautement spécialisé voire professionnel. Dans ce cadre, nous retrouvons des applications telles que :

- *teleshopping*,
- imagerie satellitaire (*Remote Sensing Applications*),
- applications bio-médicales (*Bio-medical applications*),
- édition semi-automatique de documents multi-média (*Semi-automated multimedia editing*),
- applications éducatives (*Educational applications*),
- télé-surveillance (*Surveillance applications*),
- contrôle par la vision (*Visually-based control*),
- accès Universel (*Universal Access*).

### 1.3.3 Structuration de MPEG-7

Pour optimiser l'efficacité de ses développements, MPEG est structuré de façon matricielle [MPEG-Struct] en plusieurs groupes qui rassemblent chacun des compétences spécifiques (ex. audio, vidéo) et qui visent à apporter des solutions technologiques aux requêtes et objectifs distincts et identifiés qui leur sont propres. Ces différents groupes interviennent dans les activités liées aux différents standards MPEG. En ce qui concerne MPEG-7, il bénéficie de l'apport des groupes suivants :

- *Requirements*,
- *Systems*,
- *Audio*,
- *Video*,
- *Reference Software*,
- *Conformance*,

qui sont des composantes classiques dans l'organisation des activités MPEG et auxquels s'ajoute le groupe nouveau MDS (*Multimedia Description Schemes*). Celui-ci est d'ailleurs maintenant en charge de MPEG-21. Soulignons en outre la spécificité du sous-groupe DDL (*Description Definition Language*) du groupe

*Systems*, dédié aux développements MPEG-7 sur le langage de description.

Ces divers groupes travaillent *a priori* indépendamment les uns des autres. Toutefois, pour garantir une bonne harmonisation de l'ensemble, des réunions communes aux différents groupes (ex. MDS – Video ou DDL –MDS) sont organisées lors de chaque réunion MPEG.
Soulignons le rôle central joué par le groupe *Requirements* qui possède la vision d'ensemble des objectifs et qui apparaît comme un véritable chef d'orchestre du futur standard.

Poursuivons notre présentation sur MPEG-7 par un bref aperçu de la partie MDS, qui rassemble l'ensemble des outils du futur standard et nous donne une vision globale et synthétique sur MPEG-7.

## 1.3.4 MPEG-7 MDS

Les schémas de description sont actuellement organisés en quatre couches distinctes (N3964, N3966, Mars 2001) (Figure 1.8) :

- Eléments de base,
- Gestion et description du contenu,
- Navigation et accès,
- Organisation du contenu,
- Interaction avec l'utilisateur.

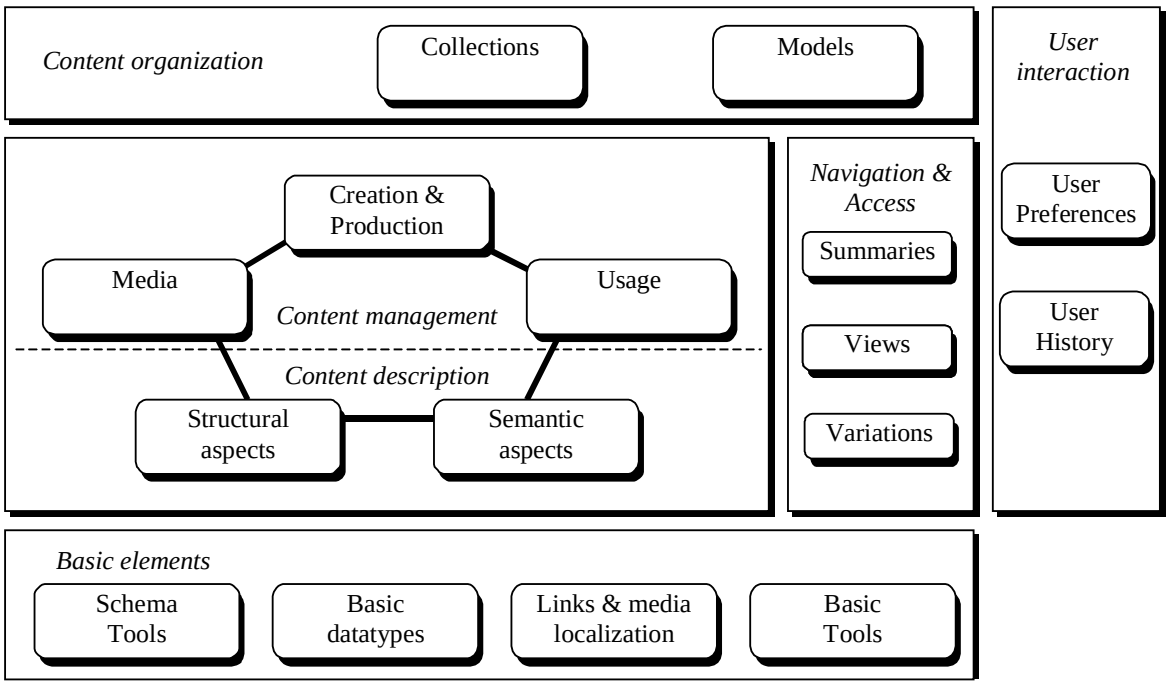



**Figure 1.8.** Couches et éléments composants de MDS [MPEG-7MDS].

La couche des éléments de base comporte :

- des outils spécifiques du DDL (*Schema Tools*),

- des structures et des types de données élémentaires (types entiers, vecteurs et matrices, histogrammes),
- des éléments de localisation du média (*link* & *media localization*) tels que *References to Ds and DSs* et *Unique Identifier*,
- des éléments temporels (*TimePoint D, Duration D, IncrDuration D, RelTime D, RelIncrTime D, Time DS, MediaTime DS*),
- des éléments de répérage des médias (*Media locators*), tels que *MediaURL D, MediaLocator DS, VideoSegmentLocator DS, ImageLocator DS, AudioSegmentLocator DS, SoundLocator DS*.

En outre, la couche des éléments de base contient des DS élémentaires, relatifs aux descriptions textuelles (*Language attribute, Language D, ControlledTerm D, Annotation DS*), des descriptions de personnes (*Individual DS, Quasiperson DS, Organization DS, Person DS*), des descriptions de situation (Place DS), des descriptions de priorités (*Weight DS*) et des descriptions des graphes de relations (*Entity-relationship graph DS*).

S'appuyant sur les éléments de base, la couche de gestion et description du contenu forme le cœur même de l'ensemble des outils de MDS. La description du contenu prend en compte aussi bien des aspects structuraux que sémantiques (conceptuels). Concernant la gestion du contenu, les éléments mis en jeu couvrent les notions de stockage des médias (par exemple format du fichier, type de codage), de droit d'usage et incluent des méta-informations essentiellement relatives à la création et à la production du contenu. Ces éléments sont brièvement explicités dans le Tableau 1.1.

| ***Ensemble d'éléments*** | ***Fonctionnalité*** |
|---|---|
| Création et Production | Métadonnées relatives à la création et la production du contenu, comme par exemple, titre du document audio-visuel (AV), auteur, classification en genres et types, objectifs. Ces éléments d'information sont ajoutés par une annotation manuelle. |
| Utilisation | Métadonnées relatives aux droits d'usage du document AV, comme les droits d'auteurs, d'accès et de publication, information financière. |
| Média | Descriptions du stockage du média, incluant le format de stockage, le type de codage du média, et d'autres éléments utiles pour l'identification des documents AV. Plusieurs types de stockage du même document AV peuvent être ici précisés. |
| Aspects structuraux | Description structurale du contenu AV, centrée sur des segments correspondant à des éléments spatiaux, temporels ou spatio-temporels distincts du contenu AV. Chaque segment contient des descripteurs adaptés à sa spécificité. |
| Aspects conceptuels | Description des éléments sémantiques du contenu AV s'appuyant sur des notions de haut niveau en termes d'objets, d'événements, de notions abstraites et de relations entre ces divers éléments. |

**Tableau 1.1.** Description des fonctionnalités de la couche *Content Management and Description.*

Un rôle à part dans cette construction revient au sous-ensemble d'éléments spécifiant les aspects structuraux du contenu (AV). En définissant de manière hiérarchique la structure spatio-temporelle d'un document AV, ce sous-ensemble constitue la colonne vertébrale de la description. En même temps, c'est la porte d'entrée qui permet l'intégration des descripteurs individuels, relatifs aux informations de forme, couleur, texture et mouvement.

Concernant la couche "Navigation et accès", des outils spécifiques incluant plusieurs types de résumés, hiérarchiques ou séquentiels, rendent possible le *browsing* (parcours rapide) des documents AV. Des modifications du contenu AV sont également spécifiées, afin d'adapter les présentations multimédias aux terminaux des clients, aux préférences de l'utilisateur, ainsi qu'aux conditions de transmission sur différents réseaux et.

La couche "Organisation du contenu" définit des éléments relatifs à la classification et la structuration des documents AV en *clusters* et collections. Cette couche regroupe fondamentalement les instruments d'analyse et de classification. Elle inclut des modèles probabilistes et analytiques, des fonctions statistiques, des structures probabilistes (ex. distribution gaussienne), des clusters et des classifieurs.

Enfin, la couche "d'interaction" avec l'utilisateur spécifie un certain nombre de SD précisant des informations permettant de créer des profils utilisateurs et de rendre possible une gestion personnalisée du contenu.

Cette brève analyse des outils MPEG-7 montre que seuls ceux relatifs aux aspects structuraux du contenu offrent des points d'entrée aux représentations par le contenu. C'est bien cette partie qui sera détaillée plus avant et exploitée dans ce travail. Les autres éléments MDS, bien qu'indispensables pour une large gamme d'applications professionnelles, sont en général fondés sur des représentations textuelles et annotations manuelles, sortant donc du contexte spécifique de l'analyse automatique d'image dont nous traitons dans cette thèse.

Dans la suite, nous proposons une revue plus approfondie des descripteurs visuels bas-niveau qui constituent les éléments porteurs d'information des représentations par le contenu. Nous verrons au Chapitre 4 comment ces différents descripteurs sont intégrés de manière synchronisée dans des schémas de description spécifiques de MDS et comment ils peuvent être exploités concrètement dans le cadre d'un système complet d'indexation incluant aussi bien des outils d'annotation, de visualisation, qu'un moteur de recherche.

### 1.3.5 Descripteurs visuels MPEG-7

Les descripteurs visuels de MPEG-7 se réfèrent aux attributs classiques et largement utilisés dans le domaine de l'indexation par le contenu, de couleur, texture, forme et mouvement. Ils incluent cinq

structures de base et vingt descripteurs. Le Tableau 1.2 résume ces différents éléments, en fonction de leurs types et des primitives d'image associées.

Les cinq structures de base sont définies comme suit :

- **Grille spatiale** (Grid *Layout*) : spécifiant la partition d'une image en blocs rectangulaires auxquels peuvent être associées des descriptions individuelles en termes de couleur, texture et mouvement.
- **Repère Spatial 2D** (*Spatial Coordinate*) : fixant un système de coordonnées 2D pour exprimer les descripteurs faisant intervenir au sein de leur définition des coordonnées, comme celui de localisation spatiale ou spatio-temporelle, de trajectoire et de mouvement paramétrique.
- **Séries Temporelles** (*Temporal Series*) : précisant des mécanismes d'échantillonnage temporel (uniforme ou non), afin d'associer à une séquence vidéo une suite de descripteurs visuels d'image fixe.
- **Interpolation Temporelle** (*Temporal Interpolation*) : proposant un mécanisme générique d'interpolation entre entités arbitraires, et intervenant principalement dans le descripteur de trajectoire.
- **Vues Multiples** (*Multiview*) : c'est en fait un SD, définissant un "*container*", qui permet d'intégrer des descripteurs. Pour l'heure *Vues Multiples* est particulièrement dédié aux descripteurs de forme 2D visant à reconnaître un objet 3D à partir d'un certain nombre de projections 2D. En effet, le descripteur de forme 2D qui est associé à chaque projection est utilisé dans la structure *Vues Multiples*.

Notons que *Vues Multiples* est la seule structure de base qui permette d'effectuer des requêtes par similarité (de forme) et qui exploite donc des mesures de similarité associées aux descripteurs individuels. Le coût de calcul est ici quadratique en fonction du nombre de vues.

| **Structures de base** | **Couleur** | **Texture** |
|---|---|---|
| Grille spatiale | Espace de couleur | Histogramme des orientations des contours |
| Série temporelle | Quantification de couleur | Texture homogène |
| Vues multiples | Histogramme de couleur scalable | Parcours rapide à partir de la texture |
| Repère spatial 2D | Couleur dominante | |
| Interpolation temporelle | Couleur d'un groupe de trames | |
| | Couleur structurée | |
| | Distribution spatiale de couleur | |

| **Forme** | **Mouvement** | **Localisation** | **Autres** |
|---|---|---|---|
| Forme - Région | Mouvement de la caméra | Localisation spatiale | Reconnaissance de visage |
| Forme - Contour | Trajectoire | Localisation spatio-temporelle | |
| Forme 3D | Mouvement paramétrique 2D | | |
| | Activité de mouvement | | |

**Tableau 1.2.** Les descripteurs visuels MPEG-7.

### 1.3.5.1 Descripteurs de couleur

Attribut d'image largement utilisé dans le domaine de l'indexation et de la représentation par le contenu, la couleur a été également considérée dans le cadre de MPEG-7. Sept descripteurs de couleur sont actuellement adoptés dans le futur standard. Nous allons en donner, pour chacun, le principe et les principales caractéristiques.

**Descripteur Espace de Couleur**
**(*Color Space Descriptor*)**

Ce descripteur est un outil auxiliaire, permettant de préciser l'espace de couleur dans lequel tous les autres descripteurs de couleur sont exprimés et de garantir l'interopérabilité indispensable entres ceux-ci.
N'étant jamais utilisé seul dans un contexte de requête par similarité, aucune mesure de similarité n'est associée à ce descripteur.
Les espaces de couleurs actuellement supportés par MPEG-7 (*cf.* Annexe 1.3) et auxquels renvoie ce descripteur sont les suivants :

- L'espace RGB, considéré comme l'espace de référence.
- L'espace $YC_bC_r$, où une couleur est représentée par une composante de luminance Y et deux différences de couleurs $C_b$ et $C_r$.
- L'espace *Monochrome*, sous-espace 1D de l'espace $YC_bC_r$ obtenu en ne conservant que la composante de luminance Y. Cet espace est particulièrement adapté lorsqu'il s'agit de considérer des images à niveaux de gris.
- L'espace *Matrice Linéaire* (*LinearMatrix)* est un espace abstrait, dont les trois composantes, notées ($C_1$, $C_2$, $C_3$), sont obtenues à partir de l'espace de référence RGB par une transformation linéaire générique, spécifiée par une matrice de transformation de dimension $(3 \times 3)$, définie par l'utilisateur. L'intérêt de cet espace est de rendre compatible MPEG-7 tout espace de couleur dérivé de RGB par une transformation linéaire.

Contrairement aux espaces de couleur précédemment mentionnés, les espaces HSV et HMMD sont dérivés de RGB par des transformations non-linéaires.

- L'espace HSV (*Hue-Saturation-Value*) tend à représenter une couleur sous forme de trois composantes correspondant aux éléments de la perception humaine. La nuance H est une mesure angulaire qui fait référence à la composition spectrale de la couleur, la valeur V donne des indications sur le niveau de luminance, tandis que la saturation S exprime la pureté de la couleur.
- L'espace HMMD (*Hue-Min-Max-Diff*) est une variante 4D de l'espace HSV, H ayant la même définition que dans le cas du HSV. La saturation S est ici remplacée par la composante Diff, définie comme la différence entre la valeur maximale (Max) des trois composantes primaires R, G et B et celle minimale (Min).

Ces différents espaces de couleur adoptés montrent que la richesse des représentations de la couleur est bien prise en compte dans MPEG-7. Notons toutefois quelques absences notamment parmi les espaces visant spécifiquement à obtenir une échelle de chromaticité uniforme comme Lab, Yuv, $U^*V^*W^*$et leurs différentes variantes [Pratt78, Jain89].

En pratique, les espaces de couleur requièrent de disposer de schémas de quantification spécifiques, pour définir différents autres descripteurs de couleur, comme par exemple ceux à base d'histogramme.

**Descripteur Quantification de Couleur**
**(*Color Quantization Descriptor*)**

Comme le descripteur espace de couleur, ce descripteur est un outil auxiliaire. Plusieurs méthodes de quantification ont été considérées dans les stades préliminaires de MPEG-7, incluant des quantifications linéaires et non-linéaires, des structures de type LUT (*Look Up Table*), offrant un support aux méthodes de quantification vectorielle... Elles avait initialement comme objectif de définir, conjointement avec les espaces de couleur, un outil flexible et configurable, permettant de spécifier un large éventail d'histogrammes de couleurs. Toutefois, aujourd'hui, au stade FDIS de MPEG-7, nous ne retrouvons qu'un outil très simple, permettant de spécifier une quantification uniforme des axes de différents espaces de couleur, en précisant notamment le nombre d'intervalles de quantification de chaque composante.
L'explication de ce changement radical d'optique du Groupe Vidéo Couleur/Texture se trouve dans les problèmes *d'interopérabilité* soulevés par une trop grande flexibilité des outils standardisés. Rappelons en effet que l'objectif primordial d'un standard est de garantir l'intéropérabilité de diverses applications exploitant les technologies normalisées. Cela demande de minimiser le nombre des technologies adoptées, selon des critères conjoints de performance, fonctionnalité et simplicité de mise en oeuvre.
Dans le cas particulier des descripteurs de couleur, comment peut-on rendre interopérables des descriptions fondées sur des espaces de couleur différents avec des quantifications également différentes ?
Un travail approfondi et impliquant de nombreux participants a été consacré à ce problème pendant plusieurs mois dans le cadre des CE Couleur/Texture [CE-CT99.12]. La conclusion des expérimentations et débats peut être résumée comme suit :

- Offrir aux utilisateurs MPEG-7 des descripteurs de couleur optimisés en termes d'espace et de quantification de couleur, plutôt que de leur laisser le libre arbitre,
- Définir, là où c'est possible, des descripteurs *scalables* au sens où différentes représentations correspondant, par exemple, à différentes quantifications, peuvent être déduites l'une de l'autre.

Suivant ces recommandations, tous les descripteurs de couleur ont été par la suite optimisés, selon des critères expérimentaux, pour un certain espace de couleur, qui sera explicitement mentionné pour chaque descripteur, et pour quelques méthodes de quantification. Quant aux aspects de scalabilité, ils ont été

essentiellement pris en compte au niveau du descripteur histogramme scalable, que nous présenterons plus loin.

Ces deux premiers descripteurs[1] sont en effet des outils élémentaires et auxiliaires, servant à définir par la suite les autres outils de description fondés sur la couleur. Poursuivons donc notre présentation par les descripteurs de couleurs offrant de véritables représentations par le contenu.

Pour des raisons historiques, commençons par le plus simple d'entre eux : l'histogramme de couleur. Même si actuellement il n'est plus un descripteur MPEG-7, étant intégré et optimisé dans le cadre du descripteur histogramme scalable, présenté plus loin, cela va permettre de fixer le concept et de définir les mesures de similarité associées.

**Descripteur Histogramme de Couleur**
**(*Color Histogram Descriptor*)**

Les histogrammes de couleur [Swain91, Stricker94, Hafner95] sont largement répandus dans le domaine de l'indexation des images par le contenu. MPEG-7 a considéré ces techniques dans ses stades préliminaires. En les utilisant conjointement avec des descripteurs comme l'espace de couleur et la quantification des couleurs, MPEG-7 a initialement adopté une approche flexible et configurable, capable de satisfaire les exigences d'une large gamme d'applications.

L'histogramme $h$ d'une image (où, plus généralement, d'une région quelconque correspondant à un objet de forme arbitraire) est défini par les fréquences relatives d'apparition des couleurs. Les couleurs de l'image sont tout d'abord quantifiées en un nombre $N$ de couleurs prototypes, notées $\{c_1, c_2, \ldots, c_N\}$, suivant les spécifications du descripteur de quantification décrit précédemment. On a alors :

$$\forall\, i \in \{1, 2, \ldots, N\}, \qquad h(i) = \frac{nombre\,de\,pixels\,de\,couleur\,c_i}{nombre\,total\,de\,pixels} \ . \tag{1.1}$$

Plusieurs mesures de similarité relatives aux histogrammes de couleur sont suggérées dans MPEG-7, depuis les plus simples, comme les distances $L_1$, $L_2$ et de Hamming (pour des histogrammes binaires), jusqu'à des mesures plus élaborées, comme la distance $L_2$ pondérée par des coefficients de similarité entre couleurs. Détaillons cette dernière.

Proposée dans [Hafner95], cette mesure est fondée sur une distance $L_2$ pondérée entre deux histogrammes, $h_A$ et $h_B$, et exprimée sous la forme quadratique suivante :

$$d(h^A, h^B) = (h^A - h^B)^\tau W (h^A - h^B), \tag{1.2}$$

[1] Les principaux contributeurs ont été IBM Watson (Etats-Unis), Heinrich Hertz Institut (Allemagne), et LGCIT (Corée).

où

- $h^A = (h_1^A, h_2^A, ..., \quad h_N^A)^\tau$, $\quad h^B = (h_1^B, h_2^B, ..., h_N^B)^\tau$,
- $N$ est le nombre d'intervalles de quantification de l'histogramme,
- $W = (w_{ij})_{i,j=1}^N$ avec $w_{ij} = 1 - \frac{\delta(c_i, c_j)}{\delta_{\max}}$, est la matrice de pondération,
- $\delta(c_i, c_j)$ est une mesure de dissimilarité entre les couleurs $c_i$ et $c_j$ ; en pratique, la mesure $\delta(c_i, c_j)$ est classiquement la distance $L_1$ ou $L_2$ ;
- $\delta_{\max}$ est la distance maximale entre deux couleurs de l'espace considéré.

On peut démontrer [Hafner95] que lorsque la mesure $\delta(c_i, c_j)$ satisfait les axiomes d'une distance, cette forme quadratique devient positive semi-définie dans le sous-espace de $R^N$ défini par :

$$\left\{ u \in R^N \quad | \sum_{i=1}^{N} u(i) = 0 \right\}. \tag{1.3}$$

Remarquons que la condition (1.3) est toujours vérifiée par les différences d'histogrammes normalisés.

Cette mesure a l'avantage de prendre mieux en compte les relations de similarité entre des couleurs proches mais correspondant à des intervalles de quantification différents. En revanche, son calcul est de complexité quadratique par rapport au nombre d'intervalles de l'histogramme.

Les analyses comparées, conduites dans le cadre des *Core Experiments* MPEG-7, ont prouvé la supériorité des représentations à base d'histogrammes dans les espaces HSV ou HMMD par rapport à celles liées à RGB, en terme de performances de requêtes par similarité. Les résultats expérimentaux ont également démontré qu'une bonne représentation du contenu de l'image en terme de couleur est obtenue avec un nombre d'intervalles d'histogramme variant entre 64 et 256.

Ces différentes optimisations ont conduit à l'intégration de l'histogramme de couleur dans une représentation scalable fondée sur la transformée de Haar.

**Histogramme Scalable par Transformée de Haar**
**(*Scalable Color Descriptor*)**

La propriété de scalabilité de l'histogramme de couleur dans l'espace HSV est exprimée à partir de la transformée de Haar [Poularakis00] qui permet de déduire des variantes de l'histogramme initial à

différentes résolutions[1] [Krishnamachari00a, Krishnamachari00b]. Le caractère hiérarchique de la transformée de Haar permet donc de comparer des histogrammes à différentes résolutions.
En outre, les coefficients sont quantifiés de manière non-linéaire en un nombre de *bitplanes*, et de manière optimisée pour l'espace HSV, ce qui conduit à un deuxième type de scalabilité, allant des représentations binaires des coefficients jusqu'aux représentations en pleine résolution.

Les mesures de similarité recommandées par MPEG-7 s'appuient aussi bien sur la distance $L_1$ dans l'espace des coefficients de Haar, que sur les mesures de similarité précédemment mentionnées, dans l'espace des histogrammes.

L'histogramme de couleur et sa version scalable s'appliquent à des images fixes. Comment prendre en compte le caractère dynamique d'une séquence vidéo ? Un réponse est apportée par le descripteur suivant.

**Descripteur Histogramme de Couleur d'un Ensemble de Trames**
**(*Group of Frames/Pictures Histogram Descriptor*)**

Ce descripteur[2] [Fermann00] étend les histogrammes de couleur précédemment définis pour des images fixes aux séquences vidéos et a comme objectif de prendre en compte la nature dynamique de celles-ci. Cette fois, l'histogramme caractérise un ensemble d'images, qui peut correspondre, par exemple, à toutes les trames d'un plan ou d'une scène.

Trois types distincts d'histogramme ont été adoptés :

- l'histogramme moyen,
- l'histogramme médian,
- l'histogramme intersection.

L'histogramme moyen est la moyenne arithmétique des histogrammes individuels sur chaque image (trame) de l'ensemble vidéo considéré. Si l'on désigne par $(h_i)_{i=1}^{N}$, les $N$ histogrammes représentés sur $N_{bins}$, alors l'histogramme moyen, noté $h_{moy}$, s'exprime par :

$$\forall\, j \in \{1, 2, \ldots, N_{bins}\}, \qquad h_{moy}(j) = \frac{1}{N}\sum_{i=1}^{N} h_i(j). \tag{1.4}$$

L'histogramme moyen décrit ainsi globalement l'information de couleur présente dans la séquence.

L'histogramme médian, noté $h_{med}$, est donné par :

[1] Descripteur promu par Philips Research (Etats-Unis) et NEC Corporation – C&C Media Research Laboratories (Japon).
[2] Descripteur promu par Philips Research (Etats-Unis) L'Université de Rochester (Etats-Unis) et Eastmann Kodak, (Etats-Unis).

$$\forall\, j \in \{1, 2, \ldots, N_{bins}\}, \qquad h_{med}(j) = \mathrm{Med}\{h_i(j)\}_{i=1}^{N}. \tag{1.5}$$

Il permet de construire une représentation synthétisant les couleurs qui sont présentes de manière majoritaire dans l'ensemble des trames considérées.

Enfin, l'histogramme intersection, noté $h_{int}$, prend la valeur minimale de chaque intervalle de quantification, conduisant ainsi à une représentation des couleurs les plus persistantes dans l'ensemble des images :

$$\forall\, j \in \{1, 2, \ldots, N_{bins}\}, \qquad h_{int}(j) = \mathrm{Min}\{h_i(j)\}_{j=1}^{N}. \tag{1.6}$$

Tous les histogrammes sont finalement re-normalisés de telle sorte que la somme de leurs éléments soit égale à l'unité.
Les mesures de similarité associées à ce descripteur sont les mêmes que celles pour les histogrammes d'images fixes.

Les descripteurs à base d'histogrammes de couleur offrent des représentations caractérisant globalement la couleur d'une image, et présentent un comportement d'invariance satisfaisant par rapport aux transformations de similarité. Toutefois, leur inconvénient majeur est lié à la perte de toute information de localisation spatiale. Pour y remédier, un nouveau descripteur prenant en compte un "minimum" d'information spatiale a été proposé et adopté par MPEG-7. Il s'agit de l'histogramme couleur-structure que nous présentons ci-dessous.

**Histogramme Couleur-Structure *(CS)***
***(Color-Structure Histogram Descriptor)***

Ce descripteur[1] étend et enrichit la notion d'histogramme, en introduisant un peu de cohérence spatiale locale. Pour ce faire, ce ne sont plus les pixels qui sont comptés et enregistrés dans les intervalles de l'histogramme, mais des *éléments structurants* (en utilisant la terminologie de la morphologie mathématique [Serra82, Serra88]). Le principe de construction l'histogramme CS, illustrée Figure 1.9, est le suivant. L'élément structurant, représenté par un masque binaire, est translaté en chaque pixel de l'image. Tous les intervalles de l'histogramme correspondant aux couleurs présentes à l'intérieur du masque sont incrémentés. Ainsi, si l'on note $h_{CS}$ l'histogramme CS, $h_{CS}(i)$ devient la fréquence relative des éléments structurants contenant la couleur d'index i.
Remarquons que l'histogramme classique est un histogramme CS, dans lequel l'élément structurant est réduit à un pixel.

Précisons qu'au niveau de MPEG-7, l'élément structurant est toujours un carré de $(8 \times 8)$ pixels. Afin de pouvoir gérer des images de taille variable et de garantir une certaine invariance par rapport aux

[1] Descripteur promu par Sharp Laboratories (Etats-Unis).

homothéties, les images sont initialement normalisées à une taille fixe approchant au mieux une image de référence de taille (320x240 pixels).

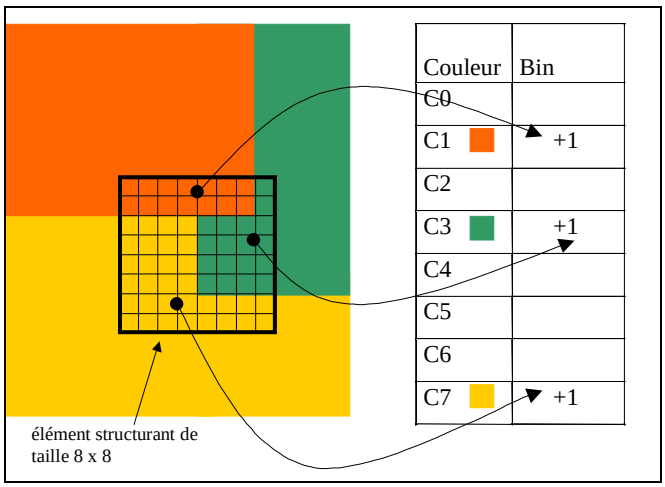
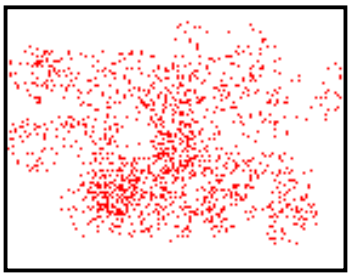
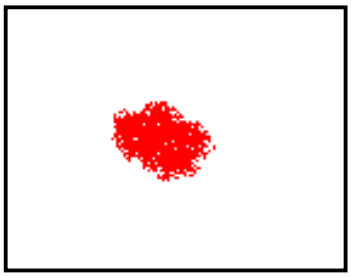



**Figure 1.9.** Principe de calcul de l'histogramme couleur-structure [MPEG-7-Visual].

Ce nouveau descripteur parvient à discriminer des informations de couleur là où les histogrammes de couleur classiques échouent, comme lorsque les couleurs sont globalement représentées de manière équivalente, mais réparties différemment dans l'image (Figure 1.10). Plus précisément, on définit la cohérence d'une couleur dominante comme le rapport entre le nombre de pixels cohérents et le nombre total de pixels dans la couleur considérée. Un pixel est considéré comme cohérent si, dans un masque rectangulaire de taille prédéfinie centré en ce pixel, il existe au moins un certain pourcentage de pixels ayant la même couleur que le pixel central.



a. Image avec cohérence de couleur réduite.



b. Image avec cohérence de couleur élevée.

**Figure 1.10.** L'histogramme couleur-structure discrimine les images a et b alors que les histogrammes classiques les confondent [MPEG-7-Visual].

Les mesures de similarité précédemment présentées s'appliquent également au cas de l'histogramme CS.

Tel que défini, l'histogramme CS est une variante simplifiée des méthodes à base de vecteurs de cohérence de couleur, proposées dans [Pass96]. Toutefois, les CE sur la couleur, conduits au sein du groupe Visuel de MPEG-7, ont montré nettement la supériorité de cette technique par rapport aux autres représentations de couleur.

Tous les descripteurs de couleur mentionnés jusqu'à présent sont fondés sur des histogrammes. Ces représentations portent en elles deux limitations principales :

- La quantification de l'espace de couleur considéré est totalement arbitraire et sans aucun rapport avec le contenu spécifique de chaque image analysée. L'histogramme contiendra par conséquent de nombreux éléments nuls.
- Les mesures de similarité à base de la distance $L_2$ pondérée (Equation (1.2)) ont une complexité quadratique avec le nombre d'intervalles des histogrammes. Elles ne sont donc efficaces que pour des représentations de dimension réduite.

Pour ces raisons, un descripteur plus adapté au contenu spécifique de chaque image a été adopté dans MPEG-7. Il s'agit du descripteur par couleurs dominantes.

**Descripteur par Couleurs Dominantes**
***(Dominant Color Descriptor)***

Son principe consiste à déterminer un nombre réduit (au maximum 8 dans MPEG-7) de couleurs principalement représentées dans l'image, appelées *couleurs dominantes*. Le descripteur[1] spécifie ensuite le nombre, les valeurs et les fréquences relatives des couleurs dominantes trouvées. Ne nécessitant pas de quantification, ce descripteur peut être utilisé dans tous les espaces de couleur MPEG-7.

Pour déterminer les couleurs dominantes, il est nécessaire d'appliquer un algorithme de segmentation (*clustering*), comme par exemple l'algorithme de quantification vectorielle LBG [Linde80] rappelé succinctement dans l'Annexe 1.4.
Une étape complémentaire d'agrégation est effectuée, afin de diminuer encore davantage le nombre final de couleurs dominantes. L'idée est de fusionner successivement les deux vecteurs les plus proches de l'ensemble de couleurs prototypes tant que la distance entre ceux-ci n'excède pas un seuil pré-défini, noté $T_d$.

Précisons qu'en ce qui concerne l'extraction, MPEG-7 recommande de représenter les couleurs dans l'espace LUV [Jain89], dans lequel la distance euclidienne entre couleurs est bien adaptée à la perception humaine de la similarité colorimétrique.

En option, le descripteur peut également intégrer les variances des classes ainsi obtenues et une mesure de confiance définie comme la moyenne des cohérences des couleurs dominantes, pondérée par leurs fréquences relatives d'apparition dans la région d'intérêt.

Les mesures de similarité proposées pour ce descripteur sont dérivées de la distance $L_2$ pondérée définie pour les histogrammes de couleur.

[1] Descripteur promu par l'Université de Californie à Santa Barbara (Etats-Unis) et par Canon Inc. (Japon).

Soient $F_1 = \left(c_{11}, c_{12}, \ldots, c_{1N_1}\right)$ et $F_2 = \left(c_{21}, c_{22}, \ldots, c_{2N_2}\right)$ deux vecteurs de couleurs dominantes et $P_1 = (p_{11}, p_{12}, \ldots, p_{1N_1})$ et $P_2 = (p_{21}, p_{22}, \ldots, p_{2N_2})$, les deux vecteurs normalisés de fréquences relatives respectivement associés :

$$\forall\, i \in \{1,2\}, \qquad \sum_{j=1}^{N_i} p_{ij} = 1. \tag{1.7}$$

Le degré de similarité entre deux couleurs dominantes $c_{1k}$ et $c_{2l}$, noté $w_{1k,2l}$, et défini par :

$$w_{1k,2l} = \begin{cases} 1 - \dfrac{\delta(c_{1k}, c_{2l})}{\delta_{\max}}, & \text{si } \delta(c_{1k}, c_{2l}) \le T_d \\ 0 & \text{sinon} \end{cases}, \tag{1.8}$$

où $\delta(c_{1k}, c_{2l})$ désigne la distance euclidienne entre deux couleurs $c_{1k}$ et $c_{2l}$.

En remarquant que, par construction, la distance entre deux couleurs dominantes de la même représentation étant toujours supérieure à $T_d$, l'équivalent de la mesure de similarité définie par l'équation (1.2), noté $D^2(F_1, F_2)$, s'exprime par :

$$D^2(F_1, F_2) = \sum_{i=1}^{N_1} p_{1i}^2 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 2 w_{1i2j}\, p_{1i}\, p_{2j}\ . \tag{1.9}$$

Une deuxième mesure de similarité, plus élaborée, est également recommandée par MPEG-7. Il s'agit ici de remplacer les poids $w_{1k2l}$ par des coefficients à base d'une distance de Mahalanobis [Mahalanobis36, Lefebvre83], exprimant la probabilité conjointe des deux couleurs dominantes, et calculés sous l'hypothèse d'une répartition gaussienne des couleurs dominantes.

**Distribution Spatiale de Couleur**
***(Color Layout - CL - Descriptor):***

Contrairement aux descripteurs de couleur mentionnés ci-dessus qui quantifient le taux global de présence de différentes couleurs, le descripteur *CL*[1] [Yamada00] capture la manière dont les couleurs sont réparties spatialement dans l'image. Cela le rend adapté aussi bien à des applications de type requête par l'exemple qu'à des requêtes par l'esquisse (*query by sketch*), où l'utilisateur, à l'aide d'une interface graphique, crée rapidement une image exprimant grossièrement la répartition spatiale des couleurs souhaitée, qu'il soumet ensuite comme exemple à un système d'indexation.

Le principe de ce descripteur est le suivant. L'image est tout d'abord divisée en 64 (8x8) blocs rectangulaires. Chaque rectangle est représenté par sa couleur dominante, qui est la moyenne (marginale) des couleurs des pixels constituant le rectangle considéré. L'espace de représentation est ici YCrCb.

On construit ainsi trois matrices 8 x 8, une par composante de couleur. Les coefficients quantifiés des transformées en cosinus discret 2D (DCT – *Discrete Cosine Transform*) de chacune de ces matrices,

[1] Descripteur promu par NEC Corporation – C&C Media Research Laboratories (Japon).

définissent le descripteur. Les coefficients sont parcourus en zigzag, à partir des coefficients basse fréquence comme illustré Figure 1.11.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 6 | 7 | 15 | 16 | 28 | 29 |
| 3 | 5 | 8 | 14 | 17 | 27 | 30 | 43 |
| 4 | 9 | 13 | 18 | 26 | 31 | 42 | 44 |
| 10 | 12 | 19 | 25 | 32 | 41 | 45 | 54 |
| 11 | 20 | 24 | 33 | 40 | 46 | 53 | 55 |
| 21 | 23 | 34 | 39 | 47 | 52 | 56 | 61 |
| 22 | 35 | 38 | 48 | 51 | 57 | 60 | 62 |
| 36 | 37 | 49 | 50 | 58 | 59 | 63 | 64 |

**Figure 1.11.** Parcours en zigzag des coefficients DCT.

Les propriétés bien connues de décorrélation et de compacité de l'énergie du signal de la DCT permettent de retenir un nombre très réduit de coefficients, aboutissant ainsi à une description très compacte. Quand ce nombre n'est pas spécifié par l'utilisateur, les valeurs par défaut sont de 6 coefficients pour la composante Y et de 3 coefficients pour chacune des composantes de couleur.

La mesure de similarité recommandée pour ce descripteur est une distance $L_2$ pondérée, exprimée par :

$$D = \sqrt{\sum_{i=1}^{N} \lambda_{Yi} (Y_i^{(1)} - Y_i^{(2)})^2} + \sqrt{\sum_{i=1}^{N} \lambda_{C_b i} (C_{bi}^{(1)} - C_{bi}^{(2)})^2} + \sqrt{\sum_{i=1}^{N} \lambda_{C_r i} (C_{ri}^{(1)} - C_{ri}^{(2)})^2}\ , \qquad (1.10)$$

où

- $N = 64$ est le nombre total de coefficients DCT,
- $\lambda_{Yi}$, $\lambda_{C_b i}$ et $\lambda_{C_r i}$ désignent respectivement les poids associés au coefficient $i$ (*cf.* numérotation des coefficients illustrée Figure 1.11) pour chaque plan de couleur $Y$, $C_b$ et $C_r$.

Le descripteur CL conclut la synthèse sur les descripteurs visuels MPEG-7 à base de couleur. Remarquons que la couleur est richement représentée au niveau du standard, sept descripteurs de couleur lui étant dédiés sur un total de vingt descripteurs visuels dans MPEG-7. Cet ensemble d'outils, en apparente contradiction avec l'un des principes immuables de MPEG qui est "*One tool, one fonctionnality*", s'articule en effet par rapport aux fonctionnalités spécifiques de chaque descripteur : outils communs de représentation de la couleur pour les descripteurs espace et quantification de couleur, requêtes globales par la couleur et scalabilité dans le cas des histogrammes à base de transformée de Haar, requêtes par l'esquisse dans le cas du descripteur CL, descriptions par *blobs* de couleurs dominantes...

Notons également que tous ces descripteurs s'appliquent aussi bien globalement, pour décrire le contenu d'une image, que localement, pour décrire la couleur des objets de forme arbitraire, préalablement identifiés par un procédé de segmentation. Quant aux relations spatiales entre ces différents éléments, elles sont

prises en compte au niveau des technologies spécifiques de MDS, où un ensemble complet de relations spatiales, temporelles et spatio-temporelles est défini.

Parmi les grands absents et donc candidats potentiels aux versions futures de MPEG-7, citons les techniques fondées sur les corrélogrammes [Huang97] ou les vecteurs de cohérence [Pass96], les représentations à base de logique floue [Tolias99, Han00, Vertan00a, Vertan00b], ou encore les modélisations par mélange gaussien [Hammoud00].

Intéressons-nous à présent aux représentations à base de texture adoptées dans le futur standard.

### 1.3.5.2 Descripteurs de texture

Dans le cadre de MPEG-7, trois descripteurs de texture ont été considérés. Commençons par le plus simple, dédié à la description texturale d'images naturelles génériques.

***Histogramme d'Orientations des Contours***
***(Luminance Edge Histogram Descriptor)***

Ce descripteur[1] représente la distribution des orientations des contours de l'image, grossièrement classifiées en 5 catégories : horizontale, verticale, diagonale à 45°, diagonale à 135° et non-directionnelle.

Pour obtenir une description plus complète, les distributions sont calculées selon trois niveaux de localisation : global, semi-global et local. Dans le premier cas, un seul vecteur de 5 éléments est généré. Il contient les fréquences relatives d'apparition de chaque type d'arête dans toute l'image. Au niveau local, l'image est divisée en 16 (4x4) blocs rectangulaires et le descripteur est calculé pour chaque bloc. Enfin, au niveau semi-global, le descripteur s'exprime sous forme de projections cumulant les distributions locales verticalement et horizontalement (Figure 1.12).

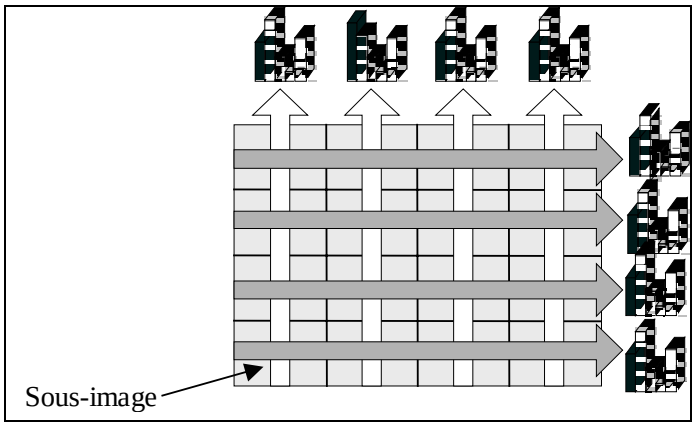
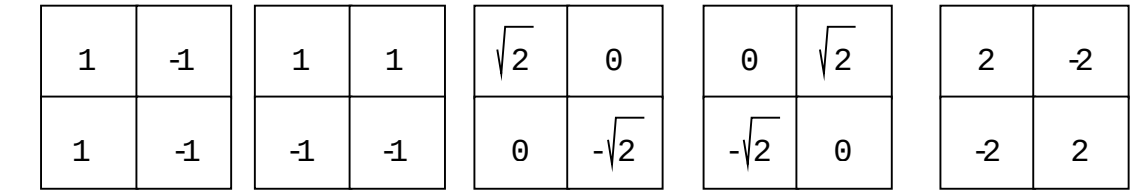


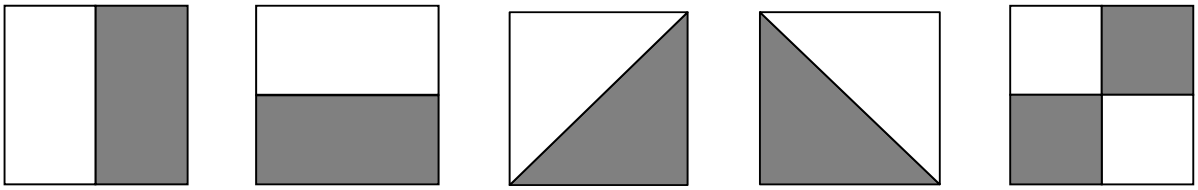

**Figure 1.12.** Calcul des distributions semi-globales d'orientations des contours [MPEG-7Visual].

[1] Descripteur promu par Ricoh Corporation (Japon), l'Université de Dongguk (Corée) et Philips Research (Etats-Unis).

Pour déterminer ces différentes distributions d'arêtes, un procédé simple est suggéré. Tout d'abord, on réalise une partition de l'image en blocs carrés. Chaque bloc est divisé en 4 quadrants sur chacun desquels est calculée la luminance moyenne. Ensuite, les opérateurs linéaires directionnels définis par les masques présentés Figure 1.13 sont appliqués à celle-ci.




a. Les filtres directionnels.



b. Les types d'arête associés.

**Figure 1.13.** Le principe de détection des différents types d'arête lors de l'extraction du descripteur histogramme d'orientation des contours.

Si la sortie maximale en valeur absolue de ces 5 filtres dépasse un seuil pré-défini, alors on décide que le bloc considéré contient une arrête du type correspondant à ce filtre.

Une simple distance $L_1$, entre les divers histogrammes obtenus, est suggérée comme mesure de similarité. Ainsi défini, le descripteur n'est pas invariant aux rotations, sauf dans le cas de rotations de faibles amplitudes, en raison de la quantification grossière (d'un pas de 90°) des directions considérées.

Les histogrammes d'orientation sont dédiés à des applications génériques de requêtes par similarité d'images naturelles. Toutefois, lorsqu'il s'agit de textures homogènes bien spécifiques, la description fournie se révèle trop élémentaire. MPEG-7 a donc adopté un deuxième descripteur de texture, proposant des représentations plus élaborées, à base d'une analyse spectrale énergétique multi-résolution.

### Descripteur de Texture Homogène (*Homogeneous Texture Descriptor*)

Le descripteur de texture homogène[1] [Manjunath96] est fondé sur des mesures énergétiques dans des sous-bandes du domaine spectral de Fourier correspondant à un banc de filtres de Gabor [Chui92].

[1] Descripteur promu par l'Université de Californie à Santa Barbara (Etats-Unis), Samsung Electronics (Corée), Heinrich Hertz Institut (Allemagne) et ETRI (Corée).

Les filtres de Gabor, bien connus dans la littérature pour leur propriété d'optimalité de la représentation temps-fréquence d'un signal (*cf.* Annexe 1.5) [Chui92, Flandrin93], sont en effet des filtres linéaires passe-bande dont la fonction de transfert est une gaussienne, définie par :

$$G(\omega,\theta;\omega_c,\theta_c,\sigma_\omega,\sigma_\theta) = e^{-\left[\frac{\sigma_\omega^2(\omega-\omega_c)^2+\sigma_\theta^2(\theta-\theta_c)^2}{2}\right]}, \tag{1.11}$$

où

- $\omega$ et $\theta$ désignent les représentations en coordonnées polaires[1] des fréquences spatiales $\omega_x$ et $\omega_y$, exprimées par :

$$\omega = \sqrt{\omega_x^2 + \omega_y^2} \qquad \text{et} \qquad \tan(\theta) = \frac{\omega_y}{\omega_x} \tag{1.12}$$

- $\omega_c, \theta_c, \sigma_\omega, \sigma_\theta$ représentent les fréquences centrales et les largeurs de bande du filtre, selon les deux coordonnées polaires contrôlant l'échelle ($\omega$) et l'orientation ($\theta$), respectivement.

A partir de cette fonction génératrice gaussienne 2D, une famille de fonctions est construite en faisant varier les paramètres d'échelle et d'orientation. Plus précisément, les sous-bandes spectrales sont définies en échantillonnant de manière dyadique le domaine des fréquences spatiales et en prenant comme fréquences centrales des filtres de Gabor :

$$\begin{aligned} &\forall\, s \in \{0,1,2,3,4\}, \qquad \omega_s = \omega_0 2^{-s\Delta}\ , \\ &\forall\, r \in \{0,1,2,3,4,5\}, \qquad \theta_r = \theta_0 + \frac{2\pi}{6} r\ . \end{aligned} \tag{1.13}$$

Quant aux dispersions, $\sigma_\theta$ est gardé constant, tandis que $\sigma_\omega$ varie de manière inversement proportionnelle à $\omega_s$.

Cette décomposition du domaine fréquentiel en 5 échelles et 6 orientations est illustrée Figure 1.14, où nous avons simulé les réponses énergétiques des 30 filtres de Gabor. Notons qu'il est suffisant de ne considérer qu'un demi-plan du domaine des fréquences, puisqu'en pratique les images texturées sont des fonctions à valeurs réelles et donc que leur spectre est symétrique par rapport à l'origine du repère considéré.

[1] Mentionnons qu'une construction similaire d'un filtre de Gabor 2D est également possible en utilisant des coordonnées cartésiennes ($\omega_x$, $\omega_y$). Néanmoins, la construction en coordonnées polaires est préférable puisqu'elle conduit à une meilleure couverture du domaine spectral.

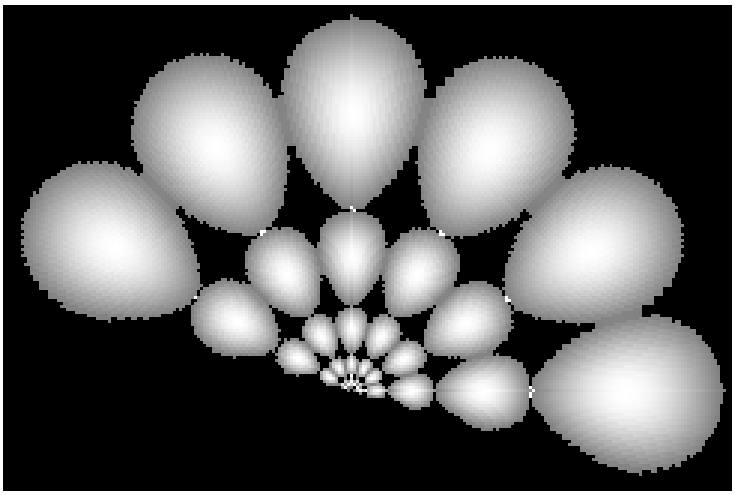


**Figure 1.14.** Réponses fréquentielles des filtres de Gabor.

Soit $S(\omega,\theta)$ la transformée de Fourier (exprimée en coordonnées polaires) de l'image à analyser. En numérotant les 30 sous-bandes selon la règle exprimée par :

$$i = 6 \times s + r + 1$$

et en notant par $G_i(\omega,\theta)$ la fonction de transfert de la sous-bande *i*, on calcule dans chacune l'énergie moyenne, notée $e_i$, de l'image et sa dispersion, $\sigma_i$ :

$$e_i = \int_{\omega}\int_{\theta} \left| G_i(\omega,\theta)\, S(\omega,\theta) \right|^2 |\omega|\, d\omega\, d\theta, \tag{1.14}$$

$$\sigma_i = \int_{\omega}\int_{\theta} \left[ \left| G_i(\omega,\theta)\, S(\omega,\theta) \right|^2 - e_i \right]^2 |\omega|\, d\omega\, d\theta. \tag{1.15}$$

Le descripteur de texture homogène est finalement défini comme le vecteur, noté *T*, ayant comme composantes les logarithmes de ces valeurs :

$$T = (\ \log(1+e_i)\ , \ldots, \log(1+\sigma_i)\ )_i\ . \tag{1.16}$$

La mesure de similarité $\delta$ s'appuie sur une distance $L_1$, pondérée par les dispersions $\alpha(k)$ de chaque élément du descripteur (calculées sur l'ensemble d'une base de textures), et est donnée ci-dessous :

$$\delta(T_1,T_2) = \sum_k \left| \frac{T_1(k) - T_2(k)}{\alpha(k)} \right|, \tag{1.17}$$

où $T_1$ et $T_2$ sont les deux descripteurs à comparer.

Tel que défini, le descripteur de texture homogène n'est pas invariant aux transformations de similarité. Afin d'assurer extrinsèquement une certaine invariance, des stratégies d'alignement sont considérées dans le calcul des mesures de similarité. Ainsi, pour gérer l'invariance aux rotations, des permutations circulaires selon le paramètre $\theta$ sont-elles générées à partir du descripteur initial. Quant aux homothéties, des

variantes à plusieurs échelles de l'image exemple sont dérivées et les descripteurs re-calculés pour chacune d'entre elles. Toutes les distances entre ces différentes variantes du descripteur et un descripteur de la base sont finalement calculées pour retenir celle qui est minimale.

Contrairement au descripteur histogramme d'orientation, ce descripteur est particulièrement utile dans le cas d'applications spécifiques de requêtes par similarité dans des bases d'images texturées uniformément.

La même analyse multiéchelle par banc de filtres de Gabor est également à la base du dernier descripteur MPEG-7 de texture, que nous présentons à présent.

**Descripteur de Parcours Rapide de Texture**
***(Texture Browsing Descriptor)***

Ce descripteur[1] [Wu00] est dédié à une classification très rapide et grossière des textures, visant comme domaine d'application le parcours rapide (*browsing*) d'une base d'images texturées. Le principe est ici de représenter une texture par des paramètres correspondant à la perception humaine, comme la direction, la régularité, l'orientation et l'échelle (la granularité). La définition de ces différents paramètres s'appuient sur la même analyse multirésolution par banc de filtres de Gabor que précédemment.

Plus précisément, les éléments retenus sont les suivants :

a. *Deux paramètres de direction* ($\theta_1$ , $\theta_2$ ), représentant les deux orientations dominantes de la texture. Ils sont issus d'une analyse par banc de filtres de Gabor, qui réalise un filtrage directionnel et mulitéchelle des textures et qui permet l'étude des variations des réponses des filtres selon les différentes orientations considérées.

b. *Deux mesures de la granularité des textures*, mesurant la périodicité des textures selon les deux directions dominantes précédemment déterminées. Cette analyse de périodicité s'appuie sur des mesures de maxima et minima locaux des signaux 1D de projections des images filtrées sur chacune des directions principales.

c. Un *paramètre de régularité*, noté $\rho$, représentant le degré de régularité (structuration) d'une texture. Cette mesure de régularité est définie à partir des autres paramètres, en étudiant le caractère persistant des projections précédemment utilisées à travers des échelles et orientations voisines.

Notons que le descripteur ainsi créé est très compact et peut être représenté sur 12 bits.

Aucune mesure de similarité ne lui est en revanche associée puisqu'il est exclusivement dédié aux applications de type parcours rapide et classification grossière des textures. Les applications associées se

[1] Descripteur promu par l'Université de Californie à Santa Barbara (Etats-Unis).

déclinent selon les 5 paramètres du descripteur et s'expriment en les termes suivants : "trouver les textures ayant une direction dominante de 30°", ou "trouver toutes les textures fortement régulières".

Mentionnons enfin, un certain nombre de travaux classiques sur l'analyse des textures, qui restent en dehors du standard, comme les représentations à base de modèles Markoviens [Li95, Graffigne98], les caractéristiques à base de matrices de co-occurrences [Haralick73, Haralick79], les travaux à base de modèles autoregressifs et de Wold [Liu96]. Il serait certainement intéressant d'en analyser les performances comparées pour enrichir la base de descripteurs texturaux de MPEG-7.

Cela conclut notre présentation des descripteurs visuels MPEG-7 de texture. Remarquons que, contrairement à la couleur, la texture est beaucoup plus modestement représentée : il s'agit ici d'uniquement trois descripteurs, dont deux seulement adaptés à des applications de requêtes par similarité. Cela s'explique en partie par la complexité et la plurivalence conceptuelle de la notion de *texture*, dont le rôle dans le monde de l'indexation générique d'images naturelles reste encore à explorer de manière plus approfondie.

Intéressons-nous à présent à un troisième attribut visuel, qui est celui de *forme*. Comment la forme d'un objet est-elle représentée au niveau du MPEG-7 ?

#### 1.3.5.3 Descripteurs de forme

La notion de forme est l'un des attributs essentiels de la perception humaine. Le large éventail de travaux consignés dans la littérature (*cf.* Chapitre 3 pour une discussion plus détaillée) essaie de répondre aux questions de modélisation, représentation et perception de forme et tente de proposer des procédures d'analyse de forme appropriées.

Au niveau du MPEG-7, les aspects de représentation de forme ont été considérés aussi bien dans le cas 2D que 3D. Introduisons tout d'abord les deux descripteurs de forme 2D actuellement retenus dans le standard. Le premier adopte une vision ensembliste, pour caractériser des formes génériques, non nécessairement représentables par un unique contour.

**Descripteur ART (*Angular Radial Transform*)**
***(Region-Based Shape Descriptor)***

Les différents objets qui apparaissent dans les images peuvent être de forme et de structure topologique assez complexe. Il n'est pas toujours possible de représenter une forme par son contour (Figure 1.15). Dans ce cas, il est préférable d'adopter une approche ensembliste : la forme est un ensemble de points du plan défini par sa fonction support. Il s'agit donc de trouver les outils appropriés pour représenter cette région support.

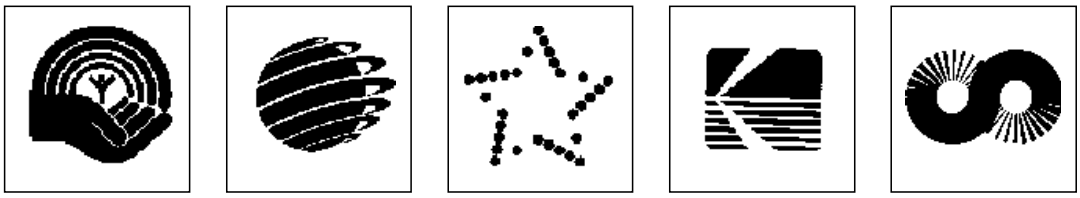


**Figure 1.15.** Exemples de formes de structure topologique complexe, non-représentables via la notion de contour [MPEG-7Visual].

Initialement, le descripteur adopté dans MPEG-7 était fondé sur une représentation de la fonction support de l'objet par moments de Zernike [Teague80, Bailey96]. Rappelons que les fonctions de Zernike sont des fonctions séparables en coordonnées polaires $(\rho,\theta)$ et qu'elles constituent un système orthonormal et complet sur l'espace des fonctions de carré sommable sur le disque unité, noté $L_2(D^2)$.

Toutefois, au cours des procédures d'évaluation MPEG-7, les promoteurs de cette technologie[1] ont substitué aux fonctions de Zernike une base de fonctions harmoniques (Figure 1.16), toujours séparables en coordonnées polaires, et définies par :

$$f_{m,n}(\rho,\theta) = \frac{1}{\pi}\cos(\pi n \rho)\, e^{jm\theta}, \tag{1.18}$$

où *m* et *n* sont des nombres entiers.
Le descripteur ART est alors défini comme l'ensemble des coefficients $c_{m,n}$ du développement de la fonction support de l'objet, notée $r(\rho,\theta)$, dans cette base, avec :

$$\forall\, m \in \{0,1,\ldots,M-1\},\ \forall\, n \in \{0,1,\ldots,N-1\}, \qquad c_{m,n} = \int_0^{2\pi}\int_0^{1} f_{m,n}(\rho,\theta)\, r(\rho,\theta)\, d\rho\, d\theta, \tag{1.19}$$

où *M* et *N* désignent le nombre de coefficients considérés (les valeurs recommandées par MPEG-7 étant de 3 et 12, respectivement).

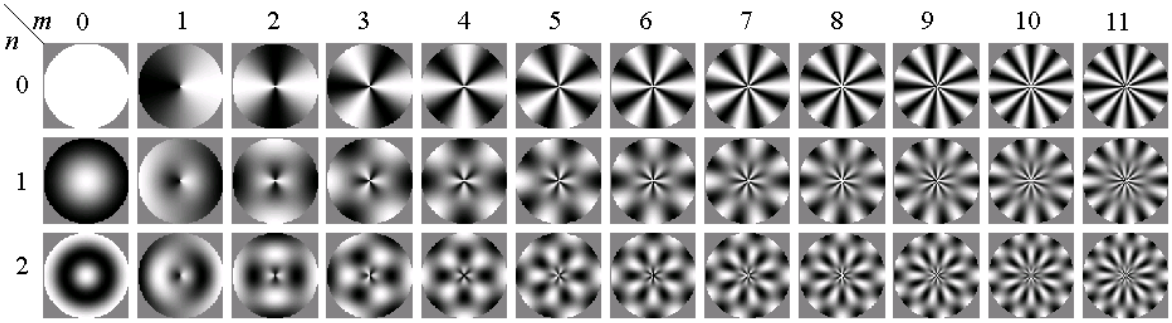



**Figure 1.16.** Les parties réelles des 36 fonctions harmoniques utilisées dans MPEG-7 [MPEG-7Visual].

Toutefois, pour pouvoir effectuer cette décomposition, la forme initiale doit être préalablement mise à l'échelle de manière à être inscrite dans le disque unité $D^2$. Cette normalisation est réalisée comme suit. Tout d'abord, l'origine du repère est placée au centre de gravité de l'objet. Ensuite, la taille de l'objet est

[1] LG Corporate Institute of Technology (Corée).

normalisée par une homothétie d'un facteur égal à l'inverse de la distance maximale entre le centre de gravité et les points de l'objet.
Cette normalisation assure une certaine invariance aux translations et homothéties. Remarquons toutefois que la normalisation de l'échelle par rapport à la distance maximale présente l'inconvénient d'une grande sensibilité au bruit et aux déformations locales de la forme.
Quant à l'invariance aux rotations, elle peut être acquise de manière intrinsèque, en considérant uniquement les valeurs absolues des coefficients de la décomposition.

Enfin, les mesures de similarité sont fondées sur la norme $L_1$ entre les vecteurs de coefficients.

Le descripteur ART présente l'avantage de la généricité, étant applicable à toute forme. Toutefois, dans des cas spécifiques où les formes considérées peuvent être représentées par un unique contour, il est plus approprié de considérer une représentation fondée sur l'information de contour, comme dans le deuxième descripteur adopté par MPEG-7 que nous décrivons par la suite.

**Descripteur par Espace Echelle de Contour (EEC)**
**(*Contour Based Shape Descriptor*)**

Ce descripteur[1] propose de décrire une forme en en caractérisant le contour. Cette approche, bien que moins générique que celle sur laquelle s'appuie le descripteur précédent, puisqu'elle ne s'applique qu'aux objets ayant des contours fermés bien définis (Figure 1.17), fournit néanmoins une représentation de forme plus élaborée.

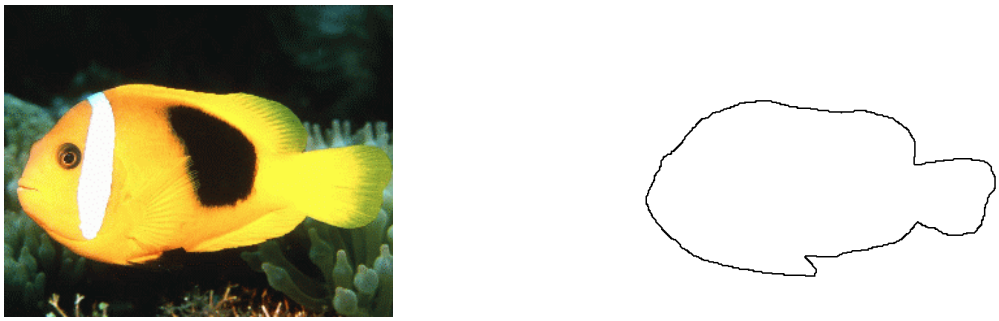


**Figure 1.17.** Un objet et son contour [MPEG-7Visual].

Le descripteur EEC exploite les maxima de courbure [Mokhtarian92], détectés à travers une analyse multi-échelle dans un espace échelle gaussien. Cette analyse multi-échelle de forme est ici nécessaire pour garantir la robustesse de la représentation, la grande sensibilité des opérateurs différentiels par rapport au bruit étant bien connue, et pour assurer la sélection des maxima de courbure prédominants, *i.e.* persistants à travers plusieurs échelles.

[1] Descripteur promu par Mitsubishi Electric Corporation (Grande Bretagne).

Considérons un repère cartésien (*xOy*). Soit ( $x(s)$, $y(s)$ )$_s$, une courbe paramétrée par l'abscisse curviligne *s*. Sa courbure, notée $k(s)$, s'exprime comme décrit ci-dessous :

$$\forall\, s \in [0,1], \qquad k(s) = \frac{x'(s)y''(s) - x''(s)y'(s)}{(x'^2(s) + y'^2(s))^{\frac{3}{2}}}, \tag{1.20}$$

où *s* est l'abscisse curviligne normalisée par la longueur totale de la courbe et les exposants ' et " désignent respectivement les dérivées du premier et du second ordre, estimées en pratique par des différences finies.

Un nombre maximal de 64 maxima de courbure est ensuite retenu, leurs coordonnées (dans le repère cartésien considéré) étant intégrées dans le corps du descripteur. En outre, deux paramètres globaux de forme, la circularité (*C*) et l'excentricité (*E*), sont ajoutés au descripteur. Ils sont respectivement définis par :

$$C = \frac{(périmètre)^2}{aire}, \qquad E = \sqrt{\frac{(m_{20} + m_{02}) + \sqrt{(m_{20} - m_{02})^2 + 4m_{11}^2}}{(m_{20} + m_{02}) - \sqrt{(m_{20} - m_{02})^2 + 4m_{11}^2}}}, \tag{1.21}$$

où $m_{ij}$ désignent les moments linéiques du second ordre du contour considéré, s'exprimant comme suit :

$$\forall\, i, j \in \{0,1,2\}, \qquad m_{ij} = \frac{1}{N}\sum_{s=1}^{N} x^i(s)y^j(s), \tag{1.22}$$

avec *N* nombre de points sur le contour.

L'invariance du descripteur par rapport aux transformations de similarité est réalisée à l'aide de mécanismes classiques d'alignement et normalisation. Ainsi, le choix de l'origine du repère au centre de gravité de la courbe garantit-il l'invariance aux translations. Pour assurer un comportement invariant aux rotations, les pics de courbure sont stockés à partir de celui de plus grande amplitude. Quant à l'invariance à l'échelle, elle est réalisée par une normalisation de type boîte englobante.

La mesure de similarité associée est fondée sur une fonction de coût d'appariement entre les maxima de courbure détectés, deux pics étant considérés comme appariés si et seulement si la distance euclidienne entre leurs positions n'excède pas un seuil prédéfini. Un terme complémentaire pénalisant les pics non-appariés est également introduit.
Les paramètres globaux de forme sont également utilisés dans l'étape d'appariement pour éliminer d'emblée les formes peu susceptibles d'être similaires à l'objet requête.

**Descripteur par Spectre de Forme 3D**
**(*3D Shape Spectrum Descriptor*)**

Le descripteur par spectre de forme 3D[1], que nous avons promu, vise à décrire la forme d'objets 3D, représentés comme des maillages 3D au format VRML.
Le SF3D est défini comme la distribution d'un index de forme caractérisant localement la géométrie d'une surface 3D et est exprimé comme la coordonnée angulaire de la représentation polaire du vecteur de courbures principales.
Ce descripteur sera présenté en détails au Chapitre 3.

### 1.3.5.4 Descripteurs de mouvement

L'information de mouvement est intrinsèquement liée au contenu spatio-temporel et dynamique des données vidéos, où plusieurs types de mouvement sont en général présents, incluant mouvements globaux de la caméra et mouvements des objets individuels.
Etudions à présent comment ces différent mouvements sont représentés dans MPEG-7.

**Descripteur de Mouvement 3D de la Caméra**
**(*Camera Motion Descriptor*)**

Ce descripteur[2] [Jeannin00] est fondé sur une modélisation complète des mouvements d'une caméra 3D, comme illustré Figure 1.18.

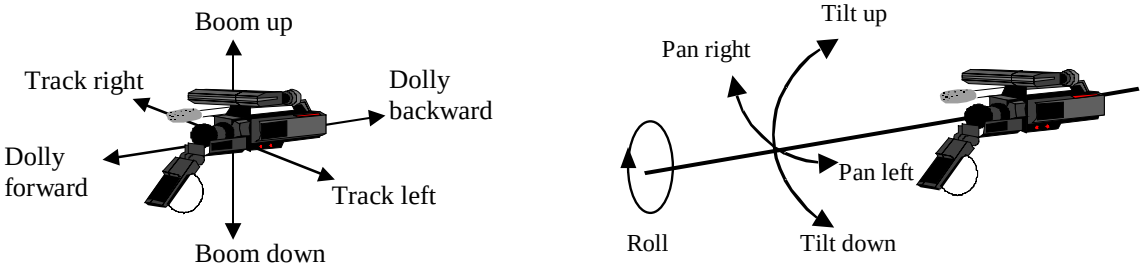



**Figure 1.18.** Les différents mouvements 3D d'une caméra [MPEG-7Visual].

Mentionnons qu'à ces mouvements correspondant à des manipulations physiques dans l'espace 3D, s'ajoutent ceux de zoom-avant et zoom-arrière, dus au changement de focale.

La Figure 1.19 présente les paramètres associés à ces différents mouvements de la caméra et le modèle de projection perspective considéré. En considérant un repère cartésien (*Oxyz*) dans l'espace 3D, le modèle de

[1] Descripteur promu par Institut National des Télécommunications (France).
[2] Descripteur promu par Laboratoires d'Electronique Philips (France).

la caméra est spécifié par trois paramètres de translation $T_x$, $T_y$, $T_z$, trois paramètres de rotation, notés $R_x$, $R_y$, $R_z$ et un paramètre qui contrôle le zoom, noté $R_{zoom}$.

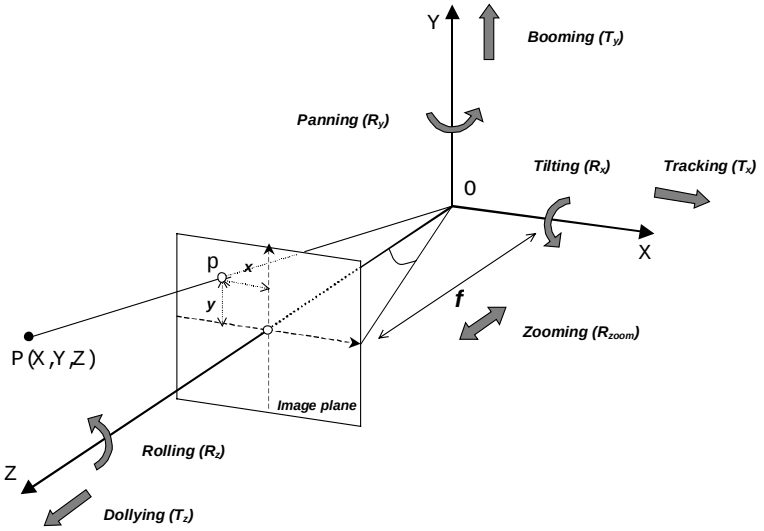



**Figure 1.19.** Les paramètres de mouvement associés à une caméra 3D et le modèle de projection perspective dans la scène 2D [MPEG-7Visual].

Les vitesses apparentes 2D, notées $(u_x, u_y)$, induites par la projection perspective des mouvements 3D de la caméra dans la scène 2D, s'expriment par :

$$u_x = -\frac{f}{Z}\left(\boldsymbol{T}_x - x\boldsymbol{T}_z\right) + \frac{x.y}{f}\cdot\boldsymbol{R}_x - f.\left(1+\frac{x^2}{f^2}\right)\cdot\boldsymbol{R}_y + y\cdot\boldsymbol{R}_z + f.\tan^{-1}\left(\frac{x}{f}\right)\left(1+\frac{x^2}{f^2}\right)\cdot\boldsymbol{R}_{zoom}\ , \tag{1.23}$$

$$u_y = -\frac{f}{Z}\left(\boldsymbol{T}_y - y\boldsymbol{T}_z\right) - \frac{x.y}{f}\cdot\boldsymbol{R}_y + f.\left(1+\frac{y^2}{f^2}\right)\cdot\boldsymbol{R}_x - x\cdot\boldsymbol{R}_z + f.\tan^{-1}\left(\frac{y}{f}\right)\left(1+\frac{y^2}{f^2}\right)\cdot\boldsymbol{R}_{zoom},\ , \tag{1.24}$$

où *f* désigne la distance focale de la caméra.

Ces équations permettent de déterminer les paramètres 3D de la caméra à partir des champs de flot optique mesurés dans la scène 2D, en utilisant des algorithmes comme ceux présentés dans [Sudhir97, Corridoni98].

Une fois ces différents paramètres déterminés, le descripteur retient les "quantités" de ces mouvements individuels, définies à base de mesures de l'aire de recouvrement d'images successives correspondant à chaque mouvement individuel de la caméra (Figure 1.20).


a. Zooming. b. Rolling

**Figure 1.20.** L'aire relative des régions en gris exprime la mesure de "quantité" de mouvement [MPEG-7Visual].

Les mesures de similarité sont définies à partir d'une distance $L_1$ pondérée par ces différentes quantités de mouvements naturels.

Le descripteur de mouvement de la caméra 3D offre ainsi une représentation déterministe du mouvement global d'une succession d'images de la vidéo.

Pour pouvoir caractériser le mouvement global sur l'ensemble des trames vidéos d'un plan ou d'une scène, une deuxième approche, à base de moments statistiques du premier et deuxième ordre, calculés sur l'ensemble des vecteurs de mouvement 2D, a été adoptée par MPEG-7. Il s'agit du descripteur d'activité de mouvement.

**Descripteur d'activité de mouvement**
***(Motion Activity Descriptor)***

Ce descripteur[1] [Divakaran98] offre une mesure de la quantité du mouvement présent dans les séquences vidéos.
Il exploite les vecteurs de compensation de mouvement associés aux *macroblocs* MPEG (blocs carrés de dimension $16 \times 16$ pixels), codés en mode prédictif. L'avantage réside ici dans le fait que ces vecteurs sont déjà inclus dans les flux MPEG-1, 2 et 4 et peuvent être déterminés sans nécessiter un décodage complet de la vidéo. Quant aux macroblocs codés en mode *intra*, qui n'ont pas de vecteur mouvement associé, leur vitesse est considérée comme nulle.

Pour chaque vecteur de vitesse $v_i = (v_{ix}, v_{iy})$, on calcule son amplitude $a_i$ et son orientation $\theta_i$, données par :

$$a_i = \sqrt{v_{ix}^2 + v_{iy}^2}\ , \qquad tg(\theta_i) = \frac{v_{iy}}{v_{ix}} \tag{1.25}$$

Puis, on estime l'amplitude et la direction moyenne, notées respectivement $\mu_a$ et $\mu_\theta$, et la variance $\sigma_a$ de l'amplitude, comme exprimé ci-dessous :

$$\mu_a = \frac{1}{N_{MB}} \sum_{i=1}^{N_{MB}} a_i\ , \qquad \sigma_a = \sqrt{\frac{1}{N_{MB}} \sum_{i=1}^{N_{MB}} (a_i - \mu_a)^2}\ , \qquad \mu_\theta = \frac{1}{N_{MB}} \sum_{i=1}^{N_{MB}} \theta_i\ , \tag{1.26}$$

où $N_{MB}$ désigne le nombre total de macroblocs sur l'ensemble des trames considérées.

Le descripteur d'activité de mouvement est composé des éléments suivants :

- *L'intensité de l'activité du mouvement*, représentée sous forme d'un nombre entier prenant ses valeurs entre 1 et 5, et obtenue par une quantification de la variance $\sigma_a$. Elle correspond intuitivement à des niveaux croissants d'activité de mouvement. Comme exemples de vidéos de

[1] Descripteur promu par Mitsubishi Electric Research Laboratories (MERL), Murray Hills (Etats-Unis).

faible activité de mouvement citons les interviews, les présentations de journaux télévisés. Parmi les séquences typiques d'une forte activité, on retrouve les retransmissions sportives ou les scènes d'action dans des films (poursuites, cascades...).

- *La direction dominante du mouvement*, résultant d'une quantification grossière de l'orientation moyenne $\mu_\theta$ sur 8 directions et entre 0 et 360 degrés.
- *La distribution temporelle de l'activité* sur l'ensemble d'un plan, représentée par un histogramme à 5 intervalles, correspondant chacun aux 5 niveaux d'intensité du mouvement. Chaque intervalle exprime la durée relative de chaque niveau d'intensité dans la séquence considérée.

En outre, afin d'obtenir des descriptions à base d'activité de mouvement à un niveau de détail plus élevé, une variante localisée spatialement du descripteur est également proposée. Elle est fondée sur un découpage de la scène vidéo en un nombre prédéfini de blocs rectangulaires.

Ce descripteur présente l'avantage de la compacité et de s'exprimer dans le domaine compressé à partir des flux MPEG. Toutefois, l'expérience montre que l'information de mouvement à base de vecteurs de mouvement obtenus par des procédures simples de compensation de mouvement (comme, par exemple, des méthodes d'appariement par blocs) est souvent peu fiable. Cela affecte de manière significative les performances du descripteur.
Le descripteur proposé est par ailleurs trop élémentaire pour prendre en compte la complexité des mouvements présents habituellement dans les vidéos naturelles.

Les deux premiers descripteurs présentés jusqu'ici visent à caractériser le mouvement global dans des scènes vidéos. Intéressons-nous à présent à la représentation du mouvement des objets individuels et commençons par le descripteur le plus simple, à base de trajectoires.

**Descripteur de Trajectoire *(Motion Trajectory Descriptor)***

Les trajectoires[1] [Jeannin00] fournissent une représentation simplifiée du mouvement d'un objet dans une séquence vidéo, à partir de points clé correspondant aux positions d'un point d'intérêt de l'objet (le plus souvent, son centre de gravité) à différents instants temporels.

Dans MPEG-7, les trajectoires considérées peuvent être aussi bien 2D que 3D. Par souci de généricité, nous traitons exclusivement du cas 3D.
Une trajectoire est représentée par un ensemble ordonné de $N$ quadruplets de nombres réels $(x_i, y_i, z_i, t_i)_{i=1..N}$, où $(x_i, y_i, z_i)$ désignent les coordonnées du point d'intérêt de l'objet dans un repère cartésien prédéfini, à l'instant $t_i$.

[1] Descripteur promu par Laboratoires d'Electronique Philips (France).

Pour déterminer les positions de l'objet à l'intérieur d'un intervalle temporel $[t_i, t_{i+1})$, un mécanisme d'interpolation est défini, s'appuyant sur une modélisation physique élémentaire, décrite par les équations suivantes :

$$\forall\ t \in [t_t, t_{i+1}),\quad \begin{cases} x(t) = x_i + v_{ix}(t-t_i) + \frac{1}{2}a_{ix}(t-t_i)^2 \\ y(t) = y_i + v_{iy}(t-t_i) + \frac{1}{2}a_{iy}(t-t_i)^2 \\ z(t) = z_i + v_{iz}(t-t_i) + \frac{1}{2}a_{iz}(t-t_i)^2 \end{cases}, \tag{1.27}$$

où $(v_{ix}, v_{iy}, v_{iz})$ et $(a_{ix}, a_{iy}, a_{iz})$ désignent respectivement le vecteur vitesse et accélération à $t_i$. Ces positions doivent satisfaire des contraintes de continuité aux bords, exprimées comme suit :

$$x(t_{i+1}) = x_{i+1}, \quad y(t_{i+1}) = y_{i+1}, \quad z(t_{i+1}) = z_{i+1}. \tag{1.28}$$

Les vecteurs accélération sont optionnels. S'ils ne sont pas présents dans les descripteurs instanciés, ils sont considérés par défaut comme nuls. En connaissant les positions spatio-temporelles et les accélérations, il est facilement possible de déduire les vecteurs vitesse à partir des conditions de continuité aux bords. Pour cette raison, les vecteurs vitesse ne sont en aucun cas retenus comme éléments du descripteur.

Les mesures de similarité s'appuient sur des distances $L_2$ pondérées en fonction des coordonnées, vitesses et accélérations.

Toutefois, le problème de l'alignement spatial, temporel et spatio-temporel des trajectoires, ainsi que celui des requêtes partielles, bien que crucial pour des applications de requêtes par similarité n'est pas abordé, malgré de nombreuses discussions [Stein98, Deng98, Prêteux-Iso99t, Caspi00, Giese00]. Cela en limite donc la portée et la pertinence.

Le descripteur de trajectoire ne peut pas prendre en compte des mouvements complexes, comme des rotations, des zooms-avant ou arrière... Pour pallier cet inconvénient, nous avons proposé et promu dans MPEG-7 un nouveau descripteur, à base d'une modélisation paramétrique 2D, que nous présentons par la suite.

### Le descripteur de mouvement paramétrique *(Parametric Motion Descriptor)*

Dans MPEG-7, le descripteur de mouvement paramétrique[1] vient compléter les descriptions élémentaires de mouvement fournies par les trajectoires, qui se révèlent être souvent insuffisantes.

[1] Descripteur promu par Institut National des Télécommunications (France), Université de Rochester (Etats-Unis) et Heinrich Hertz Institut (Allemagne).

Le principe consiste à représenter le mouvement d'un objet entre deux trames successives par un modèle de transformation géométrique paramétrique, quadratique, projective, affine, affine simplifié ou constant.
Observons en outre, que ces modèles sont associés à des objets de forme arbitraire, définis en toute généralité comme des régions spatio-temporelles (ensemble de pixels de l'image, sur un intervalle de temps), ce qui offre un cadre unifié pour représenter les mouvements d'objet et les mouvements globaux.
Les différents modèles actuellement retenus dans MPEG-7, les procédures d'extraction des paramètres et les mesures de similarité associées seront discutés en détails au Chapitre 2.

#### 1.3.5.5 Descripteurs de localisation

Les descripteurs de localisation sont des outils auxiliaires permettant de spécifier des régions spatiales ou spatio-temporelles correspondant à des objets d'intérêt dans la vidéo. Quatre descripteurs sont actuellement retenus :

- Le descripteur de localisation de région spatiale (*Region Locator*), qui définit la région soit par une boîte englobante, soit par un polygone dont les sommets sont listés dans le corps du descripteur.
- Le descripteur de localisation de région spatio-temporelle (*Spatio-temporal Locator*), qui spécifie une région sur un ensemble de trames vidéos par des séquences de descripteurs auxiliaires de trajectoire multiple et /ou de paramètres multiples. Dans le premier cas, il s'agit des trajectoires des sommets des polygones définissant la région d'intérêt. Dans le second cas, une région d'intérêt (polygone, boîte ou ellipse englobante) est suivie à l'aide des paramètres de mouvement entre les instants temporels successifs, définis à l'aide du descripteur générique d'interpolation temporelle.
- Le descripteur de trajectoire multiple (*Figure Trajectory*),
- Le descripteur de paramètre multiple (*Parameter Trajectory*).

#### 1.3.5.6 Descripteur de reconnaissance de visage

Le descripteur de visage[1] est fondé sur la technique classique de *EigenFaces* décrite dans [Turk91]. 49 vecteurs propres sont standardisés par MPEG-7 et déterminés par une analyse en composantes principales sur une base d'images de visage.

Le descripteur est défini comme le vecteur de coefficients de la décomposition sur cette base de 49 vecteurs propres.

[1] Panasonic Singapore Laboratory (Singapour), Université de New South Wales (Australie).

## 1.4 Conclusion

Fédérant les acteurs socio-économiques du monde du multimédia, le groupe MPEG de l'ISO a déjà promu les standard MPEG-1 et MPEG-2 , qui ont connu le large succès que l'on sait.
Avec le standard MPEG-4, l'horizon MPEG s'est enrichi de nouvelles fonctionnalités de représentation par le contenu, multimédia interactif et codage hybride d'objets naturels et synthétiques.
Identifiant les grands enjeux socio-économiques de disposer de technologies standardisées de description de données multimédias, le groupe MPEG a ouvert en 1998 un appel à propositions pour lancer le futur standard MPEG-7 de description de ces contenus.
Après avoir rappelé les règles, principes et procédures ISO, nous avons analysé les vingt descripteurs, standardisés MPEG-7. Ce riche ensemble d'outils de représentation par le contenu offre, d'une manière interopérable, un solide support pour un large éventail d'applications d'indexation.

# Chapitre 2

# Indexation par le mouvement des séquences vidéos


## Résumé

*Ce chapitre traite de l'indexation des documents vidéos dans le cadre de la modélisation paramétrique de mouvement. En particulier, une famille de mesures de similarité est définie dans l'espace des champs de vitesse (MSCV). Elle dépend d'une fonction distance générique spécifiée en fonction du type de requête. Les problèmes d'optimisation en temps de calcul, d'alignement spatio-temporel et de pondération des composantes translationnelle et homogène de mouvement sont analysés et une solution mathématique proposée et mise en œuvre sous forme d'une famille de mesures de similarité simplifiées exprimées dans l'espace des champs de vitesse (MSSCV).*

*Les expérimentations conduites sur des bases de test synthétique et naturelle avec vérité terrain démontrent, objectivement et quantitativement, par estimation du critère Bull Eye retenu dans le cadre de MPEG-7, la nette supériorité des MSSCV par rapport aux mesures de similarité définies dans l'espace des paramètres de mouvement et établit ainsi la pertinence du descripteur de mouvement paramétrique proposé.*



**Mots Clef**

*Indexation par le contenu, standard MPEG-7, mesures de similarité, descripteurs, modèles paramétriques de mouvement, reconnaissance de gestes, interaction homme-machine.*

## 2.1 Introduction

La problématique de l'indexation de séquences vidéos dépasse largement, par sa complexité, le cadre habituel des techniques d'indexation dédiées aux données monomédias (audio, image fixe, ...). Il s'agit ici d'indexer des contenus dynamiques, riches et hétérogènes, faisant intervenir différents types de média (image, son, texte, ...) et un très gros volume d'information.
Pour illustrer concrètement la complexité du problème, considérons une application de requête par similarité dans des bases de données. Dans le cas de données monomédias, le problème trouve une solution directe et naturelle dès qu'il est possible de définir une mesure de similarité entre les descripteurs associés aux données. Dans le cas des vidéos, il n'y a pas de solution équivalente. Comment établir si deux vidéos sont similaires ou non ? Du point de vue de la perception humaine, cette question paraît même manquer de sens. En effet, les humains ne se posent jamais le problème de la similarité globale de longues séquences vidéos, mais identifient et reconnaissent plutôt certains éléments présents dans ces vidéos, comme par exemple les acteurs d'un film, les différents événements marquants d'une transmission sportive, les paysages ou les architectures dans un documentaire.
C'est bien cette démarche par segmentation qui permet de simplifier le problème de l'indexation des documents vidéos et qui est unanimement adoptée en pratique.

Dans ce cadre générique, la vidéo est représentée par un ensemble d'éléments (ou objets) d'intérêt correspondant à des types de média bien distincts et auxquels il est possible d'associer des descripteurs élémentaires adaptés à leurs spécificités. Ce sont précisément ces éléments atomiques d'information et leurs descripteurs individuels qui seront exploités par la suite dans le cadre de diverses applications spécifiques. Les objets d'intérêt habituellement utilisés et mentionnés dans la littérature sont issus de différentes procédures de segmentation, temporelle ou spatio-temporelle. Parmi ces techniques, mentionnons les méthodes de découpage en scènes et en plans [Corridoni98, Nagasaka92, Hampapur94, Zabih95, Bouthemy99, Joly94], les procédures de segmentation de la bande sonore [Carey99, Fontaine00] et les méthodes de suivi d'objet ou de détection d'objets mobiles [Salembier99, Günsel98, Mazière00, Odobez98, Deng98, Moscheni96]. Remarquons que, dans tous les cas, s'agissant de mécanismes de segmentation, les éléments individuels doivent implicitement satisfaire certains critères de *cohérence*, préalablement définis.

Ce contexte posé, les applications d'indexation de documents vidéos se définissent ensuite par rapport aux éléments extraits. Il s'agit par exemple de retrouver, dans une vidéo ou dans une base plus large, des objets d'intérêt similaires à un exemple donné, de parcourir une vidéo selon ses éléments constitutifs et leur localisation, de créer des tables des matières ou, plus généralement, de structurer et d'organiser un contenu vidéo afin d'en rendre plus efficaces l'accès, la navigation et la visualisation. Les aspects de structuration et visualisation seront abordés en détails dans le Chapitre 4.

Quant aux représentations visuelles associées aux éléments atomiques d'information, elles exploitent les attributs classiques de forme, couleur, texture et mouvement. Les représentations fondées sur la couleur, la forme et la texture sont en général des extensions élémentaires de descripteurs 2D, comme par exemple des histogrammes de couleur moyens ou médians sur l'ensemble des trames d'un plan (*cf.* Chapitre 1

pour une synthèse des solutions adoptées par le standard MPEG-7). Dans ce chapitre, nous nous intéressons uniquement aux descriptions liées à l'information de mouvement, composante spécifique et essentielle de tout contenu spatio-temporel. Il s'agit plus précisément de résoudre les problèmes suivants :

- sélectionner les descripteurs capables de représenter efficacement le mouvement de différents objets de la scène,
- déterminer et extraire, à partir de documents vidéos, les éléments spatio-temporels satisfaisant un certain critère de cohérence par rapport à ceux-ci,
- définir des mesures de similarité appropriées, simples à mettre en œuvre et de faible complexité calculatoire, pour permettre des requêtes efficaces.

La littérature déjà riche sur ce sujet peut être synthétisée selon les 4 classes suivantes :

1. Descriptions statistiques globales,
2. Représentations du mouvement global de la caméra,
3. Représentations à base de trajectoires,
4. Représentations par modélisation paramétrique 2D.

Analysons chacune de ces catégories.

### 2.1.1 Descriptions statistiques globales

D'une façon générale, les descriptions statistiques des plans vidéos visent à caractériser globalement différents champs de vecteurs de mouvement extraits de la vidéo. Ces champs de mouvement peuvent être aussi bien des champs denses de vitesse, estimés par un algorithme de flot optique [Horn81, Lucas81] que des vecteurs de mouvement associées aux macroblocs MPEG et présents dans les flux de bits MPEG-2 et MPEG-4. Ce dernier cas correspond à ce que l'on appelle *l'indexation dans le domaine compressé* et présente l'avantage de l'efficacité, puisque les vecteurs de mouvement peuvent être déterminés sans effectuer un décodage complet de la vidéo.

Divakaran *et al.* [Divakaran98, Divakaran99] utilisent des moments statistiques du premier et deuxième ordre, calculés sur l'ensemble des vecteurs de mouvement MPEG de chaque plan pour définir des mesures telles que activité et direction dominante en terme de mouvement (*cf.* Chapitre 1 pour plus de détails). Mentionnons que cette technique a été incluse dans le futur standard MPEG-7 [MPEG-7_XM]. L'activité de mouvement est ensuite quantifiée de 1 (lente) à 5 (rapide). Cette mesure est utile pour décrire grossièrement le mouvement dans un plan et pour des applications de parcours rapide des documents vidéos.
S'appuyant toujours sur l'information des vecteurs de mouvement MPEG, Dorai et Kobla [Dorai99] déterminent tout d'abord l'orientation du mouvement dominant (quantifiée selon les 8 directions d'un voisinage en 8-connexité, V8) en projetant les vecteurs de mouvement sur chaque axe de V8 et en choisissant l'axe donnant la plus grande projection. Ensuite, en s'appuyant sur une technique fondée sur la transformée de Hough, ils détectent des mouvements particuliers de zoom et les classifient en zoom-avant et zoom-arrière. Enfin, un dernier élément relatif à l'amplitude du mouvement vient compléter la description qui aboutit à trois types de mouvement (lent, moyen et rapide).
Des descripteurs comme l'orientation du mouvement dominant et l'amplitude moyenne des vecteurs

vitesse sont également proposés dans [Zhou00].
Remarquons toutefois que de telles représentations fondées sur l'information de direction dominante du mouvement ne sont pas applicables à tous les types de mouvement naturel, comme par exemple des zooms et des rotations. Ces méthodes s'appliquent donc uniquement à des mouvements ayant une composante translationnelle dominante, mais elles n'incluent en général pas de tests dédiés pour détecter ces situations particulières (à l'exception de [Dorai99], où les zooms sont détectés séparément).

Polana et Nelson [Polana92] introduisent le concept de *texture temporelle* et proposent une analyse statistique des champs denses de vitesse mettant en œuvre des statistiques du premier ordre sur les directions et les amplitudes des vitesses (après quantification selon les 8 directions de V8) et des analyses du second ordre fondées sur des mesures définies à partir de la matrice de cooccurrence des directions des vecteurs vitesse. Cela permet d'extraire des mesures d'homogénéité sur la direction et l'amplitude du flot optique. Enfin, ces mêmes analyses statistiques sont transposées aux caractéristiques différentielles du deuxième ordre du flot optique, comme la divergence et le rotationnel.

Bouthémy et Fablet [Bouthémy98, Fablet99] utilisent également des matrices de cooccurrence mais définies cette fois dans un véritable cadre spatio-temporel 3D pour en extraire les caractéristiques de Haralick [Haralick73] et notamment celles de moyenne, variance, Dirac, moment angulaire du deuxième ordre et contraste. Les mêmes auteurs proposent ensuite dans [Fablet00-1, Fablet00-2] une modélisation markovienne des champs de vecteurs de mouvement à l'aide d'une distribution de Gibbs temporelle causale. L'estimation de cette distribution selon un critère de maximum de vraisemblance s'exprime alors comme une fonction de la matrice de cooccurrence. Les mesures de similarité définies sont fondées soit sur le test de divergence de Kullback-Liebler (approché par une procédure de Monte-Carlo pour en dériver une variante plus aisément calculable), soit sur une formalisation bayesienne utilisant un critère d*e Maximum a posteriori* (MAP). Notons que la première mesure de similarité est spécifiquement utilisée pour déduire une structuration hiérarchique arborescente des bases de séquences vidéos, après regroupement récursif des séquences individuelles dans des *clusters* calculés hors ligne. Cette structuration présente l'avantage de réduire significativement le temps de réponse aux requêtes, mais elle doit être en revanche recalculée à chaque insertion d'un nouvel élément dans la base. Remarquons de plus que ce mécanisme n'est pas applicable à des bases de données distribuées comme celles que l'on trouve actuellement sur *Internet*.

Parmi les différentes approches exploitant des histogrammes calculés sur les vecteurs de mouvement, mentionnons tout d'abord celle décrite dans [Jain98]. Les auteurs utilisent deux histogrammes, correspondant aux composantes horizontale et verticale des vecteurs de mouvement extraits des séquences vidéos par l'algorithme de flot optique de Lucas et Kanade [Lucas81]. Cette fragmentation de l'information de mouvement en deux histogrammes séparés détériore la qualité de la représentation, déjà dépourvue de toute composante de localisation spatiale.
Une approche étroitement liée à la précédente et fondée cette fois sur des histogrammes d'orientation de mouvement obtenus à partir *d'images d'histoire de mouvement* est proposée dans [Davis97, Davis99]. Plusieurs histogrammes, associés à une partition de la scène en rectangles prédéfinis, sont ici considérés afin d'enrichir la représentation en introduisant un peu d'information spatiale.

Dans la même catégorie d'approches à base d'histogrammes des vecteurs vitesse, citons la technique présentée dans [DucVo99], s'appuyant sur un histogramme 2D (en coordonnées cartésiennes) des vecteurs de déplacement MPEG. Notons que les auteurs utilisent ici leur propre méthode de calcul des vecteurs de compensation de mouvement afin de rendre l'estimation plus fiable.
Dans [Kim99], les auteurs proposent aussi des histogrammes 2D des vecteurs de compensation de mouvement, mais cette fois l'histogramme est construit en coordonnées polaires, ce qui permet une séparation intuitive des composantes en amplitude et phase.

Dans le cadre d'un système d'indexation dans le domaine compressé, parmi les divers attributs disponibles dans le flux de bits MPEG-2 (information sur les macroblocs, coefficients *DC*), Kobla *et al.* [Kobla97] exploitent l'orientation des vecteurs de mouvement associés aux macroblocs MPEG, uniformément quantifiée sur 8 bits. Les auteurs définissent une mesure de similarité de mouvement entre deux trames codées en mode prédictif comme le nombre de macroblocs correspondants ayant la même orientation de vitesse.
Une autre approche toujours fondée sur des vecteurs de mouvement MPEG est décrite dans [Sahouria98, Sahouria99]. Les auteurs mettent en oeuvre une analyse en composantes principales des champs de vecteurs pour discriminer entre les différents mouvements globaux présents dans des vidéos de retransmissions sportives. Le contenu dynamique des séquences est ensuite pris en compte soit à travers de simples moyennes temporelles, soit à l'aide d'une technique plus élaborée à base de chaînes de Markov cachées.

Les approches statistiques et celles fondées sur la modélisation des mouvements de la caméra conduisent à des descripteurs globaux fournissant une information, plus ou moins composite et fiable, du mouvement dans une scène.
Une de leurs limitations est liée à la manière de gérer la dynamique des contenus vidéos, considérant soit des successions de trames individuelles, soit des moyennes effectuées en général sur l'ensemble des trames d'un plan. Dans le premier cas, la non-stationnarité des attributs extraits est bien prise en compte, mais la description n'est plus globale. Dans le second cas, le caractère non-stationnaire du mouvement à l'intérieur d'un même plan comme typiquement dans le cas de retransmissions sportives (*cf.* Paragraphe 2.4) est complètement négligé.
Un autre inconvénient de ces approches découle du caractère souvent non-fiable des champs de vecteurs qui représentent les données d'entrée. Dans le cas des vecteurs de mouvement MPEG, ceux-ci sont en général estimés à partir de critères de compensation de mouvement définis dans un souci d'efficacité de codage des vidéos par des techniques plus ou moins optimisées de *block-matching* [MPEG-4]. Ces vecteurs ne fournissent en général pas une description fiable et fidèle du mouvement réel. En outre, ils peuvent ne pas être présents, comme dans le codage en mode *Intra* [MPEG-4]. Quant aux champs denses de vitesse extraits directement à partir des séquences vidéos, il est bien connu que l'estimation du flot optique est un problème mal posé, nécessitant l'introduction de techniques spécifiques de régularisation, complexes en temps de calcul, et en général peu fiables en présence de grands déplacements, même en les intégrant dans des schémas multi-résolution.
Enfin, de notre point de vue, ce qui condamne toutes ces approches, c'est qu'elles ne peuvent pas s'appliquer aux cas d'objets isolés quelconques.

Toujours dans un contexte de description globale du mouvement, mais cette fois dans un cadre déterministe, analysons à présent les techniques de représentation du mouvement de la caméra.

### 2.1.2 Représentations du mouvement global de la caméra

Elles reposent le plus souvent sur des modélisations paramétriques 2D ou 3D des mouvements de la caméra. Ces modèles ont le mérite de prendre en compte de manière fiable, compacte et précise une large classe de mouvements rencontrés dans les vidéos naturelles. En outre, cette modélisation paramétrique permet de créer des mosaïques, *i.e.* des images panoramiques obtenues après avoir aligné, par juxtaposition et composition, plusieurs images du même plan. Elles synthétisent le contenu de ce plan. Les modèles paramétriques 2D et les méthodes de création de mosaïques ont été largement exploités et retenus au niveau des instances internationales de normalisation, aussi bien dans le cadre du standard MPEG-4, pour des applications de codage vidéo, que dans le cadre du futur standard MPEG-7, où un schéma de description mosaïque a été spécifiquement adopté. Au-delà d'objectifs de visualisation, l'image mosaïque peut être enrichie de descripteurs spécifiques aux images fixes. Cela conduit alors à l'extension naturelle et intuitive des descripteurs d'image aux plans vidéos.

Dans [Smolic99], des modèles paramétriques 2D sont proposés pour estimer le mouvement global de la caméra, créer des mosaïques, segmenter et indexer des séquences vidéos. D'autres approches exclusivement dédiées à la création de mosaïques, aussi bien 2D que 3D, sont décrites dans [Davis98, Pope97, Kumar95].

Dans [Sudhir97], les auteurs proposent une modélisation complète de la caméra dans l'espace 3D et réalisent une décomposition des mouvements apparents induits dans les images 2D en deux composantes distinctes : une *singulière* (si le mouvement de la caméra induit un point de vitesse nulle dans le flot optique 2D) et l'autre *non-singulière* (si le flot optique associé n'a pas de point critique). Chaque type de mouvement est déterminé à partir du flot optique extrait des images et catégorisé en *pan*, *tilt* et translation verticale et horizontale (pour les mouvements non-singuliers) et translation / rotation autour de l'axe *z* (*cf.* Chapitre 1, Figure 1.19) et zoom-avant ou arrière (pour les mouvements singuliers). Une approche similaire est présentée dans [Corridoni98], où les auteurs proposent une modélisation initiale affine du mouvement de la caméra, suivie d'une classification du mouvement global en *pan*, *tilt*, zoom et *dolly*, effectuée par la technique décrite dans [Adiv85].

Dans [Jeannin00], une modélisation complète des actions de la caméra dans l'espace 3D est également proposée. Sous l'hypothèse d'un modèle de projection perspective, les contributions moyennes de chaque mouvement individuel sur l'ensemble des trames de chaque plan sont ensuite estimées (*cf.* Chapitre 1), à partir des champs de vecteurs extraits par un algorithme de flot optique. Rappelons que cette technique a été adoptée dans le standard MPEG-7.

Dans [Bouthemy00, Gelgon98], une modélisation affine du mouvement dominant permet également de classifier le mouvement en *pan*/*tilt*, zoom avant/arrière, rotation gauche/droite. La classification est fondée sur une décision Bayesienne et exploite la décomposition du mouvement affine en termes de translation, divergence, rotationnel et composante hyperbolique. Les mêmes auteurs considèrent

également une modélisation affine du mouvement pour extraire des mosaïques, segmenter et suivre des objets vidéos. Pour des objectifs de visualisation rapide du contenu vidéo, ils proposent ensuite de construire une "trame synoptique", *i.e.* l'image mosaïque avec les contours et les trajectoires des objets individuels superposés. Une approche similaire et ayant les mêmes objectifs de visualisation synoptique des scènes vidéos est décrite dans [Irani97, Pope98]. Le problème des requêtes par similarité de mouvement n'est en revanche pas abordé. Dans [Bouthemy00, Bouthemy99], les mêmes modèles de mouvement affine sont utilisés dans le cadre d'un algorithme de détection de changement de plans. La méthode s'appuie sur des mesures associées aux régions support du mouvement dominant, déterminé en appliquant l'algorithme d'estimation robuste du mouvement affine proposé dans [Odobez95].

Enfin, une approche moins conventionnelle est proposée dans [Bruno00]. S'appuyant sur une décomposition en série de Fourier des composantes du flot optique, les auteurs utilisent ces coefficients pour indexer les séquences vidéos. A partir d'une mesure de similarité fondée sur une distance dans l'espace des coefficients de Fourier, une reconnaissance des activités humaines simples, comme mouvements vers la gauche, la droite, le haut, le bas, avancer, reculer, est mise en oeuvre.

Ces représentations globales sont bien utiles pour caractériser ou catégoriser le mouvement d'une caméra et ont l'avantage, dans le cas des plans exhibant un mouvement panoramique, d'aboutir à la création de mosaïques, exploitables pour des objectifs de visualisation ou d'extraction d'autres descripteurs spécifiques aux images fixes. En outre, les mécanismes d'extraction de paramètres sous-jacents permettent de créer des outils pour des objectifs complémentaires de segmentation temporelle ou de détection et segmentation d'objets mobiles. Cependant leur utilisation directe pour des applications de requête par similarité reste peu exploitée, principalement en raison de l'absence de mesures de similarité appropriées. Nous verrons dans les paragraphes suivants, que la définition de telles mesures entre les modèles paramétriques, comme celles que nous avons introduites au cours des développements du futur standard MPEG-7 [Prêteux-Iso99.07m, Zaharia01], rend opérationnelles aussi bien la mise en œuvre des applications de requête que la gestion du contenu dynamique des séquences vidéos.

Poursuivons avec les approches dédiées à la représentation des mouvements d'objets vidéos individuels de forme quelconque. Ces objets sont spécifiés par leurs supports spatio-temporels préalablement déterminés par des algorithmes dédiés de segmentation ou de suivi d'objet, comme ceux cités au début de ce paragraphe.

Discutons tout d'abord des approches à base de trajectoire.

### 2.1.3 Représentations à base de trajectoires

Les représentations à base de trajectoire d'objet vidéo visent à fournir une description simplifiée mais intuitive du mouvement des différents composants d'une vidéo. D'une manière générale, une trajectoire est définie comme un ensemble de points $(x_t, y_t)_{t=0}^{T}$, indexés par le temps *t*, et représentant les positions d'un certain point d'intérêt dans l'objet, à des instants successifs. Le plus souvent, ce point d'intérêt est le

centre de gravité de l'objet. Il s'agit donc de déterminer comment représenter efficacement et utiliser ces trajectoires pour des applications de reconnaissance et de requêtes par similarité de mouvement.

Avant de synthétiser les approches à bases de trajectoire de la littérature, précisons tout d'abord quelques points d'ordre général. Dans le contexte générique de l'indexation par le mouvement, les approches à base de trajectoire doivent nécessairement satisfaire les contraintes habituelles d'invariance par rapport aux transformations géométriques comme translations et mise à l'échelle (spatiale, temporelle ou spatio-temporelle). Une contrainte complémentaire et spécifique aux trajectoires est liée à la définition des supports des requêtes partielles, puisque les objets suivis dans les séquences vidéos peuvent avoir des durées de vie très variables. Cela est encore plus important lorsqu'il s'agit d'application de reconnaissance de gestes, comme l'identification de différents signes dans une conversation en langue des signes entre déficients auditifs. Remarquons également que ces diverses contraintes peuvent être allégées dans le cas d'applications spécifiques, comme la télésurveillance, où le fond de la scène est supposé constant et où les coordonnées des objets dans la scène sont supposées absolues.

Comment ces diverses contraintes sont-elles prises en compte par les approches décrites dans la littérature ?

Traitant d'applications de télésurveillance [Sahouria97], les auteurs construisent tout d'abord deux vecteurs 1D, correspondant aux deux coordonnées horizontale et verticale. Ces deux vecteurs sont ensuite normalisés, aussi bien en amplitude (par division par la plus grande composante) qu'en dimension (par sous- ou sur-échantillonnage) pour obtenir des trajectoires ayant le même nombre de points. Si la normalisation en amplitude assure l'invariance à l'échelle, la possibilité d'effectuer des requêtes partielles est complètement perdue. Combinée à la normalisation en dimension, cela conduit à la perte complète de l'information de vitesse. Les auteurs y remédient en introduisant un élément complémentaire qui est la vitesse moyenne, définie comme le déplacement total sur le nombre de trames. Des variantes simplifiées des deux vecteurs sont ensuite obtenues en leur appliquant la transformée en ondelettes de Haar [Daubechies92, Chui92, Truchetet98, Mallat99, Poularakis00] et en gardant uniquement les 8 coefficients basse résolution. Des mesures de similarité fondées sur la distance euclidienne de $\mathbb{R}^{16}$ sont définies pour permettre les applications de requêtes par similarité.

Une approche plus élaborée et mathématiquement bien fondée est détaillée dans [Dağtaş00]. Les auteurs identifient tout d'abord les diverses contraintes d'invariance spatiale et de requêtes partielles à satisfaire ou non en fonction des applications envisagées et proposent des mécanismes spécifiques d'appariement de sous-séquences. Deux modèles distincts sont détaillés, l'un purement à base de trajectoires et l'autre exploitant également les boîtes englobantes des objets pour générer des *traces* (images binaires obtenues par la superposition des boîtes englobantes au cours de la séquence). Afin de définir des mesures de similarité entre trajectoires, les auteurs considèrent des techniques de recherche exhaustive du meilleur appariement. Plus précisément, si $T_1$ et $T_2$ sont deux trajectoires à comparer, respectivement composées de $N_1$ et $N_2$ points (avec $N_1 \leq N_2$), la mesure de similarité $D$ entre les deux trajectoires est définie par la relation suivante :

$$D(T_1, T_2) = \min_{\tau}\left\{\delta(T_1^{\Delta\tau}, T_2)\right\}, \tag{2.1}$$

où $T_1^{\Delta\tau}$ désigne la trajectoire $T_1$ translatée temporellement d'un nombre $\Delta\tau$ de trames, et $\delta$ une distance entre les trajectoires.
Cette approche a le mérite de pouvoir supporter des requêtes partielles, mais elle est pénalisée par sa grande complexité de calcul (quadratique par rapport au nombre de points des trajectoires). Pour cette raison, les auteurs proposent différentes méthodes pour réduire le temps de calcul afférent, comme par exemple, celle à base de convolutions calculées dans l'espace des coefficients de Fourier ou celle exploitant un schéma de filtrage des trajectoires candidates pour ne retenir que les plus similaires à la requête. Notons que l'approche fondée sur l'équation (2.1) est également proposée dans [Li97], mais sans analyse complémentaire pour en réduire le temps de calcul.

Le problème de l'appariement des trajectoires est également abordé dans [Cheng00]. Ici, les auteurs considèrent des outils spécifiques au domaine de l'appariement de chaînes de caractères, et proposent des fonctions d'association triangulaires [Wang90]. Ces mécanismes permettent de formaliser de manière unifiée les requêtes partielles et totales. Toutefois, l'alignement proposé est très coûteux en temps de calcul ($O(N^3)$ pour comparer deux trajectoires ayant $N$ points).

Dans [Jeannin00], les trajectoires sont représentées sous forme d'un ensemble de *points clé*. Des mécanismes d'interpolation (linéaire ou quadratique), fondées sur une modélisation physique du deuxième degré faisant intervenir les vitesses et accélérations (*cf.* Chapitre 1 pour une discussion plus approfondie) sont ensuite définis pour permettre la reconstruction de la trajectoire à partir de ces points clé. Rappelons que cette technique a été adoptée au sein du standard MPEG-7. Comme mesure de similarité, les auteurs proposent des distances quadratiques pondérées selon les positions (coordonnées), vitesses et accélérations des points clé. Dans le cas des trajectoires ayant des nombres de points clé différents, la trajectoire la plus longue est tronquée au plus petit nombre de points. Le problème de l'alignement spatial, temporel et spatio-temporel des trajectoires, ainsi que celui de requêtes partielles, bien que crucial pour des applications de requêtes par similarité ne sont pas abordés et aucun résultat expérimental n'est présenté dans l'article. Une approche similaire est proposée dans [You01], avec des résultats expérimentaux peu concluants obtenus sur une quarantaine de trajectoires extraites de la base de test des descripteurs de mouvement MPEG-7 et correspondant à des séquences de télésurveillance.

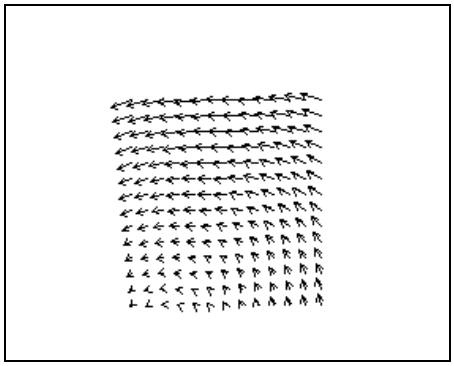
Une approche analogue et présentant les mêmes limitations dues à la distance euclidienne définie entre les coordonnées des points des trajectoires, est incluse dans le système d'indexation Video-Q [Chang98]. Ici, les trajectoires peuvent être définies par l'utilisateur à l'aide d'une interface graphique, conduisant à des requêtes par esquisse (*sketch*). Les performances du système en général et des requêtes par similarité de trajectoire en particulier, peuvent être évaluées en ligne sur le site Web [Video-Q].

Quelques approches plus sophistiquées à base de trajectoire sont traitées dans le domaine de la reconnaissance de gestes. Les applications spécifiques ciblées ici permettent d'intégrer une forte connaissance *a priori* et des règles statistiques pour décrire les différents gestes, et de prendre en compte

la nature intrinsèquement dynamique des trajectoires. Les différentes approches reposent soit sur des chaînes de Markov cachées (CMC) [Starner96], soit sur des réseaux de neurones [Yang99] pour l'apprentissage de la structure spatio-temporelle des gestes. Une approche originale et étroitement liée à celle à base de CMC proposée dans [Black98] est tout d'abord utilisée pour la reconnaissance de gestes simples et étendue ensuite à la reconnaissance des expressions faciales.

Les approches fondées sur les trajectoires caractérisent globalement le mouvement d'un objet, mais ne peuvent pas prendre en compte des mouvements plus complexes, comme ceux provenant de la combinaison de plusieurs mouvements élémentaires ou de la projection 2D de mouvements dans une scène 3D. Ces limitations sont illustrées Figure 2.1. Le train et le ballon ont la même trajectoire. Du point de vue d'un descripteur de trajectoire ces deux mouvements sont donc identiques. Toutefois, le mouvement du ballon contient une composante rotationnelle supplémentaire, ignorée lorsque l'on considère uniquement les trajectoires.
En outre, soulignons les difficiles aspects d'alignement spatio-temporel et de requêtes partielles engendrés par les méthodes à base de trajectoire, qui sont typiquement des problèmes mal-posés [Caspi00, Giese00, Stein98] et qui restent des questions ouvertes.

a. Deux trames successives de la séquence "Mobile".



b. Le flot optique dense dans la région sélectionnée
(le rectangle en blanc dans a) montre un mouvement de rotation du ballon.

**Figure 2.1.** Limitation du descripteur de trajectoire dans le cas de mouvements complexes : le ballon et le train ont la même trajectoire mais des mouvements différents.

En résumé, aucune des trois premières catégories d'approche ne permet de prendre en compte efficacement la complexité des mouvements habituellement présents dans les séquences vidéos naturelles.

Elles offrent au mieux des approximations plus ou moins grossières du mouvement réel et présentent de nombreuses limitations liées aux problèmes d'alignement, à l'estimation non-fidèle et à leur complexité... Examinons à présent les perspectives offertes par la modélisation de mouvement d'objets vidéos quelconques par transformations paramétriques 2D.

### 2.1.4 Représentations par modélisation paramétrique 2D

L'approche orientée objet, exploitant des modèles de mouvement paramétrique 2D, offre un contexte mathématique unifié pour caractériser de façon efficace et compacte le mouvement d'objets vidéos arbitraires (fond ou objets en premier plan). Les modèles paramétriques de mouvement ont été largement utilisés en estimation de mouvement [Odobez95, Dufaux93], segmentation et suivi par le mouvement d'objets vidéos [Odobez98, Deng98, Smolic99]. Au niveau du groupe MPEG, ces mêmes modèles ont été considérés dans le cadre de la norme MPEG-4 pour estimer le mouvement global de la caméra et pour créer des mosaïques [MPEG-4, MPEG-7_XM].

Dans le cadre du système d'indexation Netra-V [Deng98], un modèle affine de mouvement est mis en oeuvre pour suivre les objets vidéos et les paramètres affines sont exploités pour indexer par le mouvement de ces mêmes objets. Pourtant, les paramètres de mouvement ne sont pas utilisés pour des applications de requête par similarité. Les auteurs soulignent l'absence d'une mesure appropriée de similarité entre les mouvements paramétriques et rejettent les distances entre vecteurs de paramètres, dépourvues selon eux de sens physique.

Dans [Zaharia-Iso99.02, Prêteux-Iso99.07m, Zaharia01], les auteurs proposent une approche similaire pour caractériser le mouvement d'objets définis comme des régions spatiales décrivant un mouvement *cohérent* aussi bien spatialement que temporellement. Le mouvement de l'objet est ensuite caractérisé par le vecteur de paramètres d'un modèle perspectif planaire correspondant à une succession de trames considérée comme représentative de toute la séquence.

Reconsidérés dans un cadre hiérarchique [Ekin00], les modèles paramétriques de mouvement sont à nouveau utilisés pour décrire le mouvement dominant des objets vidéos ayant un mouvement cohérent. Une description de plus haut niveau sémantique est ensuite obtenue en regroupant les mouvements dominants élémentaires des objets en *actions* (différents mouvements dominants du même objet) et *interactions* (groupes de mouvements dominants de plusieurs objets). Les mesures de similarité entre mouvements élémentaires relèvent de distances exprimées soit dans l'espace des paramètres, soit sur les champs de vecteurs vitesse associés aux modèles paramétriques [Prêteux-Iso99.07m]. D'une façon générale, les résultats présentés ne sont pas concluants, en raison : (1) de la dimension trop petite de la base de mouvements élémentaires considérés (28 séquences seulement) et (2) de la nature articulée des mouvements considérés, mal décrits par un unique modèle de mouvement, qui contredit dans ce cas l'hypothèse de cohérence spatiale des divers mouvements.

Toutefois, malgré la compacité et la fidélité de ces représentations, leur exploitation directe dans les systèmes d'indexation existants et plus particulièrement dans le cadre des applications de requêtes par

similarité du mouvement est restée peu exploré. Deux difficultés majeures sont à mentionner :

1. le manque d'une mesure de similarité appropriée,
2. le manque de mécanismes capables de gérer correctement le caractère dynamique de l'information du mouvement dans de longues séquences vidéos.

Au cours des procédures d'évaluation des technologies soumises dans le cadre de l'élaboration de la future norme MPEG-7 et en cohérence avec ses spécifications et objectifs, nous avons démontré, en collaboration avec les équipes du Heinrich Hertz Institut (Allemagne) et du Rochester Institut (Etats-Unis), la pertinence de disposer d'un descripteur de mouvement paramétrique d'objet (DMPO) [Prêteux-Iso99.07m, Zaharia-Iso99c, Zaharia-Iso99d, Smolic-Iso99a, Smolic-Iso99b, Ekin-Iso99].

Actuellement retenu par MPEG-7, le DMPO est tout d'abord décrit puis sa mise en œuvre dans le cadre d'application de recherche et de sélection dans des bases vidéos analysée en détails. Méthodes d'extraction des paramètres (robustes ou non-robustes), influence du choix des modèles et complexité sont discutées de manière approfondie. A la lumière des verrous technologiques identifiés, nous construisons une famille de mesures de similarité définies dans l'espace des champs de vitesse (MSCV) et dépendant d'une fonction distance générique que nous spécifions en fonction du type de requête. Les problèmes d'alignement spatio-temporel et de pondération des composantes translationnelle et homogène de mouvement sont présentés et une solution mathématique proposée et mise en œuvre sous forme d'une famille de mesures de similarité simplifiées exprimées dans l'espace des champs de vitesse (MSSCV), et optimisées en terme de complexité de calcul.

Les expérimentations conduites sur des bases de test synthétique et naturelle, avec vérité terrain, que nous avons créées spécifiquement pour évaluer les descripteurs MPEG-7 de mouvement [Zaharia-Iso99e], démontrent objectivement et quantitativement par estimation du critère *Bull-Eye* retenu dans le cadre de MPEG-7, la nette supériorité des MSSCV par rapport aux mesures de similarité définies dans l'espace des paramètres de mouvement et établit ainsi la pertinence du DMPO.

## 2.2 Le descripteur de mouvement paramétrique d'objet

Présentons donc le descripteur MPEG-7 par mouvement paramétrique, en commençant par le principe de base et en allant jusqu'à sa définition en DDL MPEG-7, tel qu'actuellement inclus dans le FDIS MPEG-7.

### 2.2.1 Principe et définitions

Le principe du DMPO consiste à représenter le mouvement d'un objet vidéo entre deux trames successives sous forme d'une transformation géométrique 2D (Figure 2.2). Des séquences considérées comme *cohérentes* du point de vue du mouvement, aussi bien spatial que temporel, sont caractérisées par les paramètres de mouvement d'une transition de trames considérée comme représentative de la séquence entière. Le DMPO est ensuite défini comme le vecteur de paramètres associé à la transformation.

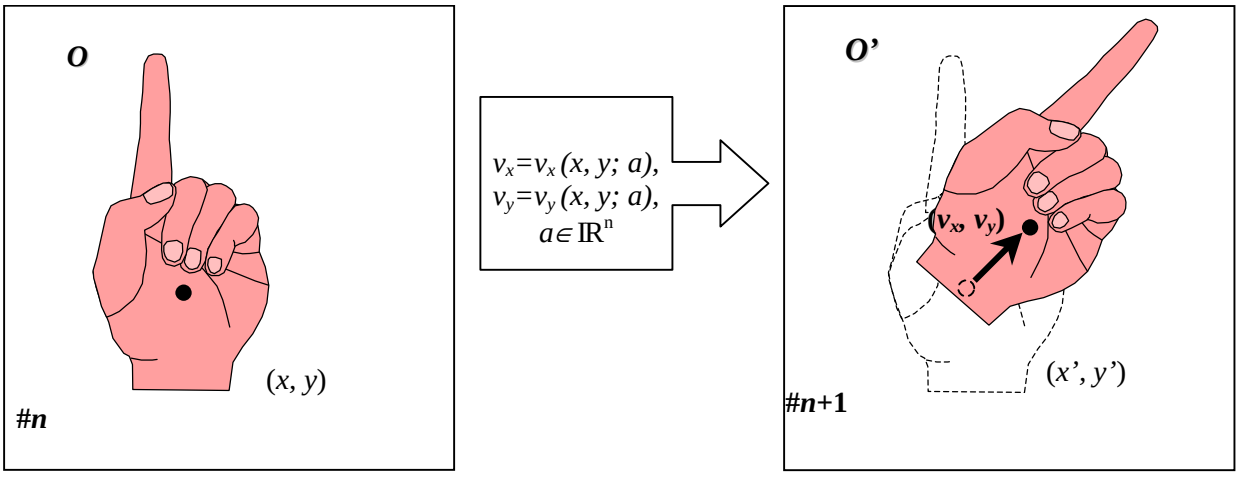



**Figure 2.2.** Principe de la modélisation paramétrique de mouvement. Le changement de la position de l'objet entre deux trames successives est représenté sous forme d'une transformation géométrique paramétrée par un vecteur de nombre réels, noté *a*.

Ce principe simple peut être formalisé comme suit. Soit un repère cartésien (*xOy*). Soient (*x*, *y*) (respectivement (*x'*, *y'*)) les coordonnées d'un pixel de l'objet dans la trame *t* (respectivement *t*+1). Notons $\Re$ la région support de l'objet considéré à la trame *t*. Pour tout point $(x, y)$ de $\Re$, les nouvelles coordonnées (*x'*, *y'*) sont données par la relation (2.2) :

$$\begin{cases} x' = x + v_x \\ y' = y + v_y \end{cases}, \quad \text{avec} \quad \begin{cases} v_x = v_x(x, y; a) \\ v_y = v_y(x, y; a) \end{cases}, \tag{2.2}$$

où $a \in \mathbb{R}^n$ est un vecteur de paramètres associé à la transformation géométrique considérée et $(v_x(x, y; a), v_y(x, y; a))_{(x,y)\in\Re}$ désigne le champ de vitesses associé à l'objet $\Re$ et au modèle de paramètres *a*.

Les performances de cette modélisation paramétrique en terme d'adéquation par rapport aux champs de vitesses apparentes réelles dépendent principalement de la complexité des modèles considérés. Le Tableau 2.1 regroupe les différents modèles que nous utiliserons par la suite et qui sont en outre actuellement retenus dans MPEG-7.

| Modèle translationnel (constant) : | Modèle affine : |
|---|---|
| $\begin{cases} v_x(x, y) = a_1 \\ v_y(x, y) = a_2 \end{cases}$ | $\begin{cases} v_x(x, y) = a_1 + a_3\, x + a_4\, y \\ v_y(x, y) = a_2 + a_5\, x + a_6\, y \end{cases}$ |
| Modèle affine simplifié (rotation/homothétie) : | Modèle perspectif planaire : |
| $\begin{cases} v_x(x, y) = a_1 + a_3\, x + a_4\, y \\ v_y(x, y) = a_2 - a_4\, x + a_3\, y \end{cases}$ | $\begin{cases} v_x(x, y) = (a_1 + a_3\, x + a_4\, y) / (1 + a_7\, x + a_8\, y) \\ v_y(x, y) = (a_2 + a_5\, x + a_6\, y) / (1 + a_7\, x + a_8\, y) \end{cases}$ |
| Modèle parabolique (quadratique) : | |
| $\begin{cases} v_x(x, y) = a_1 + a_3\, x + a_4\, y + a_7\, xy + a_9\, x^2 + a_{10}\, y^2 \\ v_y(x, y) = a_2 + a_5\, x + a_6\, y + a_8\, xy + a_{11}\, x^2 + a_{12}\, y^2 \end{cases}$ | |

**Tableau 2.1.** Les modèles paramétriques adoptés dans MPEG-7.

Les modèles affines permettent de caractériser une large classe de mouvements 2D, comme des translations, rotations, homothéties et leurs combinaisons. Les modèles perspectifs planaires et quadratiques prennent en compte des mouvements plus complexes, comme des déformations globales dues à la projection dans le plan de l'image d'objets en mouvement dans la scène 3D. Les performances relatives de ces différents modèles de mouvement seront présentées en détails au Paragraphe 2.3.5. Notons que les trois premiers modèles sont des cas particuliers du modèle perspectif ou quadratique. Néanmoins, par souci de compacité des descriptions, ces modèles sont considérés de manière indépendante dans MPEG-7, avec un champ complémentaire dans la syntaxe du descripteur pour préciser le type de modèle utilisé. De plus, des outils spécifiques d'estimation sont associés à chaque type de modèle.
Une représentation très compacte du mouvement de l'objet est ainsi obtenue. Le nombre de paramètres est de 2, dans le cas constant, de 6 pour un modèle affine et de 12 pour un modèle quadratique.

La syntaxe *XML Schema* du DMPO, telle que retenue dans le standard au stade FDIS est présentée Figure 2.3. Outre les paramètres de mouvement, représentés comme un vecteur de nombre réels, le descripteur peut inclure soit une référence vers un repère spatial, soit une définition complète de repère. Un élément supplémentaire, *MediaDuration*, spécifie la longueur de l'intervalle temporel auquel les paramètres de mouvement s'appliquent. Le champ *motionModel* spécifie le type de modèle de mouvement utilisé.

```
<complexType name=”ParametricMotionType” final=”#all”>
   <complexContent>
      <extension base=”mpeg7:VisualDType”>
         <sequence>
            <choice>
               <element name=”CoordRef”>
                  <complexType>
                     <attribute name=”ref” type=”IDREF”
                        use=”required”/>
                     <attribute name=”spatialRef” type=”boolean”
                        use=”required”/>
                  </complexType>
               </element>
               <element name=”CoordDef”>
                  <complexType>
                     <attribute name=”originX” type=”float”
                        use=”required”/>
                     <attribute name=”originY” type=”float”
                        use=”required”/>
                  </complexType>
               </element>
            </choice>
            <element name=”MediaDuration”
                  type=”mpeg7:MediaIncrDurationType”/>
            <element name=”Parameters”>
                  <simpleType>
                     <restriction base=”mpeg7:floatVector”>
                        <maxLength value=”12”/>
                     </restriction>
                  </simpleType>
            </element>
         </sequence>
         <attribute name=”motionModel” use=”required”>
            <simpleType>
               <restriction base=”string”>
                  <enumeration value=”translational”/>
                  <enumeration value=”rotationOrScaling”/>
                  <enumeration value=”affine”/>
                  <enumeration value=”perspective”/>
                  <enumeration value=”quadratic”/>
               </restriction>
            </simpleType>
         </attribute>
      </extension>
   </complexContent>
</complexType>
```

**Figure 2.3.** Syntaxe du descripteur de mouvement paramétrique exprimée conformément au langage de description MPEG-7.

Concluons la présentation du descripteur par un bref aperçu des différentes soumissions au WG11, en soulignant les principales contributions que nous avons apportées pendant le processus de standardisation.

### 2.2.2 Contexte normatif MPEG-7

Le descripteur de mouvement paramétrique a été initialement proposé [Zaharia-Iso99.02] en réponse au *Call for Proposals* (CfP) MPEG-7 [MPEG-7CfP] et examiné par un collège international d'experts lors de la réunion d'évaluation des propositions de Février 1999 (Lancaster, UK). Le descripteur a alors suscité peu d'enthousiasme parmi les évaluateurs, en raison en partie du manque de résultats sur une base de mouvements issus de la base de test MPEG-7 [MPEG-7Content], précédemment constituée et devant servir de base d'évaluation aux différents descripteurs proposés (*cf.* le document de spécification des CfP MPEG-7 [MPEG-7Eval]). En outre, l'absence de ses différents promoteurs lors de la réunion WG11 de Mars 1999 n'a pas permis l'ouverture immédiate d'un *Core Experiment* sur le mouvement paramétrique.

Une discussion scientifique et technique entre les différents supporters du mouvement paramétrique (INT, Université de Rochester - UR, Heinrich Hertz Institut - HHI) a alors eu lieu et s'est concrétisée sous forme d'une contribution commune [Smolic-Iso99a], demandant l'ouverture d'un *Core Experiment* pour évaluer le descripteur de mouvement paramétrique. En parallèle, nous avons créé la base de mouvements synthétiques [Zaharia-Iso99e], qui nous a permis de présenter de premiers résultats de requêtes par similarité de mouvement lors de la réunion MPEG de Vancouver (Juillet, 1999), fondés sur des mesures de similarité entre champs de vitesse [Prêteux-Iso99.07m]. A la lumière de ces différentes contributions, auxquelles s'ajoute la prise en compte des outils déjà présents dans le standard MPEG-4, le groupe Vidéo a accepté de promouvoir directement le descripteur dans le standard (à cette époque au simple stade de *Experimentation Model* - XM) et décidé d'ouvrir un CE pour tester les différentes mesures de similarité et recommander certaines d'entre elles dans l'ensemble des outils non-normatifs du XM.

Les mesures de similarité à base de champs de vitesse ont fait l'objet d'évaluations croisées avec implantations indépendantes. Les résultats obtenus sur la base des mouvements synthétiques, offerte par l'INT à l'ensemble des participants au CE, ont été présentés [Zaharia-Iso99c] lors de la réunion WG11 d'Octobre 1999. Le groupe Vidéo a demandé à disposer de résultats sur des mouvements naturels et a recommandé de continuer le CE. L'INT a alors créé la base des mouvements naturels de main et l'a mise à la disposition des participants. Les résultats présentés lors de la réunion WG11 de Décembre 1999 [Zaharia-Iso99d, Ekin-Iso99, Smolic-Iso99b] ont convaincu définitivement le groupe Vidéo qui a retenu deux mesures de similarité comme outils non-normatifs dans le XM MPEG-7, l'une à base de champs de vitesse et l'autre à base de distances dans l'espace des paramètres (en raison de sa simplicité).

L'intégration du descripteur dans le logiciel de référence a été réalisée conjointement par INT et HHI. Les deux équipes assurent depuis la maintenance et les différentes mises à jour nécessitées par l'évolution naturelle du standard.

Afin de rendre opérationnelle cette modélisation paramétrique de mouvement, il est toutefois nécessaire de disposer de techniques d'estimation fiables, robustes et de complexité de calcul raisonnable. Le paragraphe suivant spécifie les méthodes d'estimation que nous avons utilisées dans ce travail. Elles sont fondées principalement sur les techniques classiques spécifiées dans [Odobez95] et [MPEG-7_XM] qui recourent à une minimisation itérative multirésolution d'une fonctionnelle de l'erreur de compensation,

formalisée dans le cadre des M-estimateurs robustes [Hubert81].

Par la suite, nous considérons un repère lié à l'objet, dont l'origine est située au centre de gravité de la région support. Ainsi, les termes libres dans la définition des modèles paramétriques correspondent-ils aux composantes de translation du centre de gravité de l'objet. Un mouvement est dit *purement homogène* lorsque la vitesse du centre de gravité de l'objet est nulle.

## 2.3 Estimation des paramètres de mouvement

Les techniques d'estimation des paramètres de mouvement sont fondées sur des algorithmes de minimisation d'une fonctionnelle d'énergie de l'erreur de compensation, intégrant ou non des estimateurs robustes, ou des approches multi-résolution (voir [Stiller99] pour un *overview* sur l'estimation de mouvement, paramétrique ou non). Ce type d'approche est étroitement lié aux méthodes classiques de flot optique fondé sur l'hypothèse de photométrie constante.

Une première famille [Ekin00, Wang94] consiste à :

- Estimer le flot optique dense dans la région d'intérêt considérée en appliquant des algorithmes classiques [Horn81, Lucas81],
- Ajuster un modèle paramétrique (par exemple, en utilisant une méthode de moindres carrés) sur le champ de vecteurs vitesse précédemment calculé.

Les inconvénients de ce type d'approche viennent principalement de l'étape d'estimation du flot optique, souvent peu fiable et coûteuse en temps de calcul. Pour cette raison, nous préférons retenir la deuxième famille d'approches, qui consiste à estimer directement les paramètres de mouvement à partir des images.

Le paragraphe suivant décrit la technique d'estimation adoptée, implantée et utilisée dans ce travail, qui est fondée sur deux méthodes classiques de la littérature. La première a été introduite dans [Odobez95, Sawhney95] et consiste à résoudre le problème de minimisation par une série d'itérations de type Gauss-Newton. La seconde, introduite dans [Dufaux96, Dufaux00], adoptée dans le standard MPEG-4 initialement dans le seul cas du modèle perspectif planaire, a été ensuite étendue dans le cadre du futur standard MPEG-7 à tout autre type de modèle de mouvement du Tableau 2.1. Elle met en œuvre l'algorithme de Levenberg-Marquardt. Comme nous le verrons par la suite, même si les deux méthodes ont des formulations initiales légèrement différentes, par le biais des approximations effectuées, elles aboutissent à une solution commune, modulo le choix de la norme robuste retenue.

Présentons donc ci-dessous le principe de l'algorithme utilisé, en commençant par formaliser le problème dans le cadre spécifique des M-estimateurs robustes.

### 2.3.1 Fonctionnelles d'énergie et M-estimateurs robustes

Avec les notations introduites au Paragraphe 2.2, l'estimation de mouvement est formulée comme un problème de minimisation, dans l'espace des paramètres, de l'erreur de compensation de mouvement,

notée $E_c(a)$ :

$$E_c(a) = \sum_{(x,y)\in\Re} \rho[\, e_t(x,y;a)\,], \tag{2.3}$$

où

- $a$ représente le vecteur des paramètres,
- $e_t(x,y;a)$ désigne l'erreur de compensation au pixel $(x,y)$, à l'instant $t$, donnée par :

$$e_t(x,y;a) = I_t^c(x,y;a) - I_{t+1}(x,y)\,, \tag{2.4}$$

avec la fonction de luminance compensée $I_t^c(x,y;a)$ s'exprimant sous la forme :

$$I_t^c(x,y;a) = I_t(x + v_x(x,y;a), y + v_y(x,y;a)) \tag{2.5}$$

- $\Re$ désigne le support spatial de la région d'intérêt considérée à l'instant $t$,
- $I_t$ et $I_{t+1}$ les luminances des images aux instants $t$ et $t+1$, respectivement, et
- $\rho$ une fonction distance à préciser.

La norme $\rho$ contrôle dans l'estimation totale, l'influence des pixels présentant des erreurs importantes de compensation. D'une manière générale, la fonction $\rho$ doit être positive, puisqu'elle spécifie une distance. De plus, si des algorithmes de minimisation faisant intervenir des gradients sont utilisés, il est nécessaire d'imposer des contraintes complémentaires sur $\rho$, comme celle de dérivabilité. Le choix le plus simple satisfaisant ces conditions correspond à la norme $L_2$ et consiste à fixer $\rho(x) = x^2$. Cela revient à utiliser l'erreur quadratique de compensation comme critère d'évaluation de l'estimation. Cependant, la norme $L_2$ présente l'inconvénient d'être fortement biaisée par les erreurs individuelles de grande amplitude qui peuvent correspondre soit à du bruit, soit à des occultations, soit à des déformations non-rigides. Pour pallier cet inconvénient, d'autres fonctions peuvent être considérées, comme par exemple l'estimateur de Tuckey (suivant l'approche proposée dans [Odobez95]) ou l'estimateur $L_2$-limité utilisé dans [MPEG4, MPEG-7_XM].

La dérivée de la fonction distance est appelée *fonction d'influence* et désignée par $\psi(e) = \rho'(e)$.
On peut démontrer que la minimisation itérative par une méthode des moindre carrés itérés de l'expression (2.3) revient à considérer une norme $L_2$ pondérée, exprimée par la relation (2.6) :

$$E_c^{L_2-pondérée}(a) = \sum_{(x,y)\in\Re} w_t(x,y)\, e_t^2(x,y;a)\,, \tag{2.6}$$

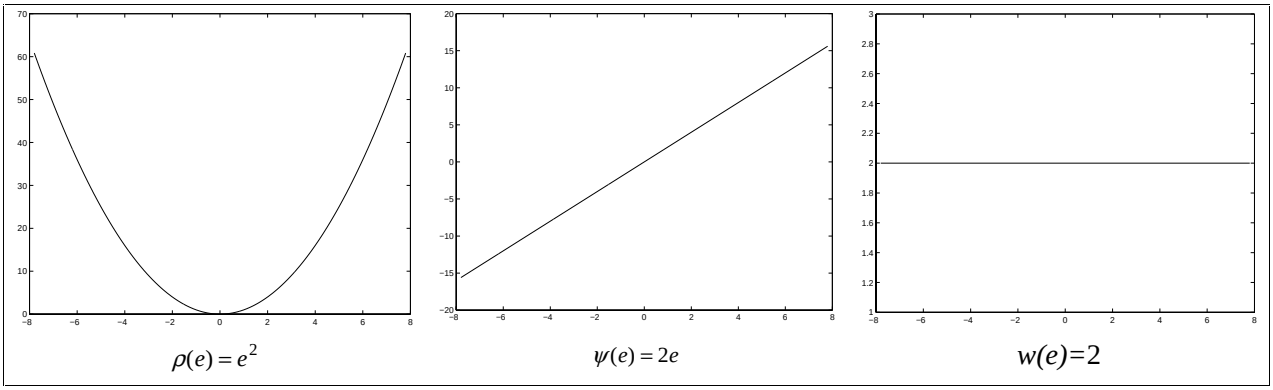
où $w_t(x,y) = w(e_t(x,y;a))$ est appelée fonction de poids, donnée par la relation (2.7) :

$$w_t(x,y) = \frac{\rho'(e_t(x,y))}{e_t(x,y)}\,. \tag{2.7}$$

La fonction ρ est choisie de telle manière que la fonction de poids associée, $w_t$, contrôle la contribution des différents pixels à l'estimation totale, adaptativement aux erreurs individuelles d'estimation $e_t(x,y)$.

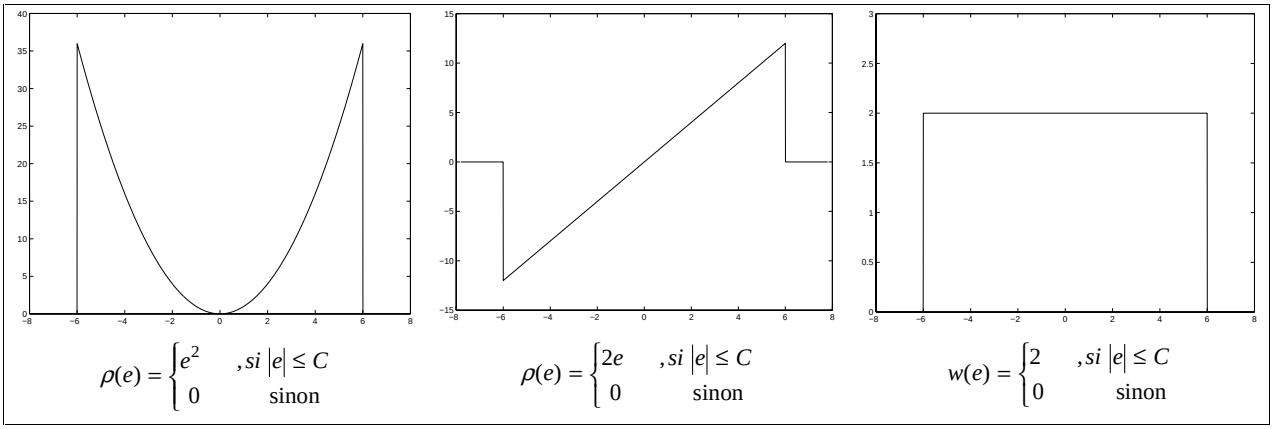
Ainsi, les pixels (*x*,*y*) pour lesquels $e_t(x,y;a) \approx 0$ seront considérés comme fidèles au modèle estimé et auront une contribution plus importante, tandis que les pixels ayant une erreur d'estimation élevée se verront attribuer un poids tendant vers 0, étant considérés comme des exceptions dues au bruit ou aux occultations.

La Figure 2.4 présente les graphiques, les relations de définition, les fonctions d'influence et les fonctions de poids des différentes fonctions distance utilisées ici.



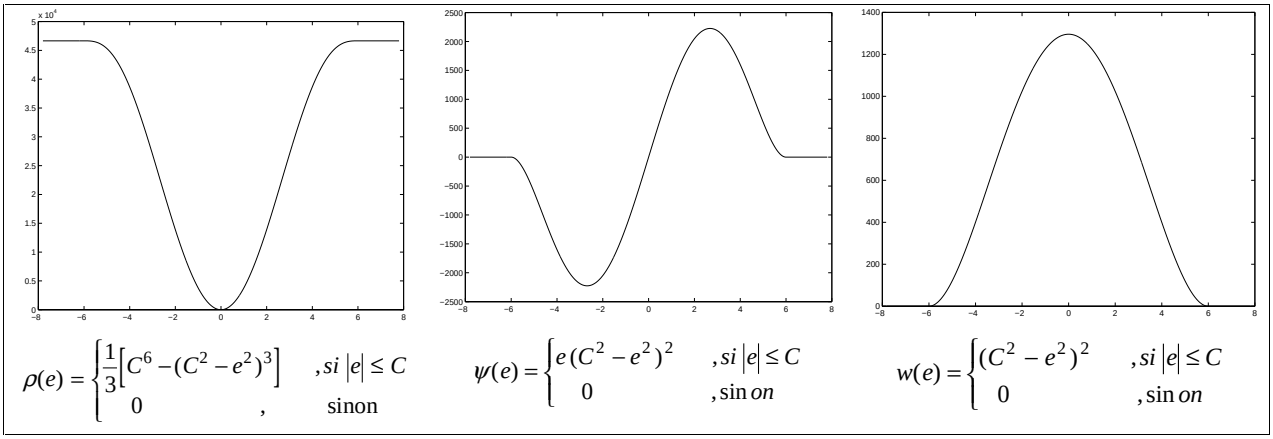
$\rho(e) = e^2$ $\psi(e) = 2e$ $w(e)=2$

a. Estimateur $L_2$.



$$\rho(e) = \begin{cases} e^2 & ,si\ |e| \le C \\ 0 & \text{sinon} \end{cases} \qquad \rho(e) = \begin{cases} 2e & ,si\ |e| \le C \\ 0 & \text{sinon} \end{cases} \qquad w(e) = \begin{cases} 2 & ,si\ |e| \le C \\ 0 & \text{sinon} \end{cases}$$

b. Estimateur L2-limité.



$$\rho(e) = \begin{cases} \frac{1}{3}\left[C^6 - (C^2 - e^2)^3\right] & ,si\ |e| \le C \\ 0 & ,\ \text{sinon} \end{cases} \qquad \psi(e) = \begin{cases} e(C^2 - e^2)^2 & ,si\ |e| \le C \\ 0 & ,\text{sin}\, on \end{cases} \qquad w(e) = \begin{cases} (C^2 - e^2)^2 & ,si\ |e| \le C \\ 0 & ,\text{sin}\, on \end{cases}$$

c. Estimateur de Tuckey.

**Figure 2.4.** Fonctions distance $\rho$, fonctions d'influence $\psi$ et fonctions de poids *w* pour différents estimateurs.

Cette formulation générique permet d'intégrer aisément tous les types de modèles présentés Paragraphe 2.2. En outre, en fonction du choix de la région support $\Re$ et du modèle utilisé, il est possible de considérer différentes applications dans un cadre mathématique unifié. Par exemple, pour une fenêtre d'estimation $\Re$ carrée et centrée successivement en chaque pixel de l'image et pour un modèle constant, on retrouve la formulation du problème du flot optique. Lorsque la région support $\Re$ est le support de l'image et que des modèles d'ordre supérieur sont utilisés, on retombe sur une estimation du mouvement global de la caméra. Enfin, lorsque $\Re$ correspond au support spatial d'un objet d'intérêt dans la scène issu d'un algorithme de segmentation, on obtient une modélisation du mouvement individuel de l'objet considéré.

Les différentes méthodes de la littérature se distinguent en fonction des algorithmes de minimisation utilisés, du choix de la norme $\rho$ et des techniques spécifiques d'estimation des gradients spatio-temporels.

Il s'agit donc de déterminer le vecteur de paramètres $\hat{a}$ minimisant l'énergie $E_c(a)$ :

$$\hat{a} = \arg\min_{a}\{E_c(a)\}. \tag{2.8}$$

Ce problème typique d'optimisation non-linéaire nécessite pour le résoudre de mettre en œuvre des algorithmes numériques dédiés. Toutefois, moyennant quelques approximations supplémentaires, il peut être simplifié et ramené à la résolution d'un système d'équations linéaires, comme nous le verrons au paragraphe suivant.

### 2.3.2 Solution optimale

Pour des vitesses $v_x$ et $v_y$ suffisamment petites, le développement de la fonction $I_t(x,y)$ en série de Taylor au voisinage du point $(x,y)$ considéré s'exprime comme suit (en négligeant les termes d'ordre supérieur à 2) :

$$I_t(x + v_x(x,y;a), y + v_y(x,y;a)) \approx I_t(x,y) + \nabla I_t^{\tau}(x,y)\, v(x,y;a), \tag{2.9}$$

où $\nabla I_t^{\tau}(x,y) = \left( \frac{\partial I_t(x,y)}{\partial x} \quad \frac{\partial I_t(x,y)}{\partial y} \right)$ désigne le vecteur gradient de l'image au point $(x,y)$.

L'expression de l'erreur de compensation s'exprime alors linéairement en fonction des vitesses $v_x$ et $v_y$ :

$$e_t(x,y;a) = \frac{\partial I_t(x,y)}{\partial t} + \nabla I_t^{\tau}(x,y)\, v(x,y;a), \tag{2.10}$$

où $\frac{\partial I_t(x,y)}{\partial t} = I_t(x,y) - I_{t+1}(x,y)$ représente la dérivée partielle par rapport au temps $t$ de l'image $I_t$ en $(x,y)$, exprimée à l'aide de différences finies.

Sous la contrainte de photométrie constante, $e_t(x,y;a) = 0$, on obtient *l'équation du flot optique*, donnée par la relation bien connue (2.11) :

$$\frac{\partial I_t(x,y)}{\partial t} + \nabla I_t^{\tau}(x,y)\, v(x,y;a) = 0\,. \tag{2.11}$$

La condition nécessaire d'extrémalité de l'énergie $E_c(a)$ s'exprime sous la forme :

$$\nabla_a E_c(a) = 0\,, \tag{2.12}$$

où $\nabla_a = \left( \frac{\partial}{\partial a_1}, \frac{\partial}{\partial a_2}, \ldots , \frac{\partial}{\partial a_P} \right)^{\tau}$ désigne l'opérateur de gradient par rapport au vecteur de paramètres $a$.

En explicitant l'équation (2.12), on obtient :

$$\nabla_a E_c(a) = \sum_{(x,y)\in\Re} \rho'(e_t(x,y;a))\, \nabla_a e_t(x,y;a) = 0\,. \tag{2.13}$$

Notons par $w_t(x,y) = \frac{\rho'(e_t(x,y;a))}{e_t(x,y;a)}$ (le rapport étant pris au sens de la limite lorsque $e_t(x,y;a) = 0$).

On obtient alors la relation suivante :

$$\nabla_a E_c(a) = \sum_{(x,y)\in\Re} w_t(x,y)\, e_t(x,y;a)\, \nabla_a e_t(x,y;a) = 0\,, \tag{2.14}$$

qui démontre l'équivalence du problème de minimisation de la fonctionnelle donnée par l'équation (2.3) avec celle de la fonctionnelle fondée sur une distance $L_2$ pondérée avec les poids $w_t(x,y)$ selon la formule ci-dessous :

$$E_c^{L_2-pondérée}(a) = \sum_{(x,y)\in\Re} w_t(x,y)\; e_t^2(x,y;a)\;, \tag{2.15}$$

Ici, la dépendance des poids $w_t(x,y)$ du vecteur de paramètres $a$ est négligée.

Tenant compte de la relation (2.10), la condition d'extremum local (2.14) devient :

$$\sum_{(x,y)\in\Re} w_t(x,y)\, \nabla I_t^{\tau}(x,y)\, v(x,y;a)\, \nabla_a I_t^c(x,y;a) = -\sum_{(x,y)\in\Re} w_t(x,y)\, \frac{\partial I_t(x,y)}{\partial t}\, \nabla_a I_t^c(x,y;a)\,. \tag{2.16}$$

Pour tous les modèles de mouvement présentés Paragraphe 2.2, à l'exception du modèle perspectif planaire, le vecteur vitesse $v$ dépend linéairement du paramètre $a$, ce qui s'exprime sous la forme du produit matriciel suivant :

$$v(x,y;a) = M(x,y)\, a\,, \tag{2.17}$$

où $M(x,y)$ est une matrice de taille $(2 \times P)$, dépendant du modèle considéré, $P$ étant le nombre de composantes du vecteur de paramètres $a$.

Notons que le modèle perspectif planaire peut être lui aussi approché par une fonction dépendant linéairement du vecteur des paramètres, en considérant un développement en série limité.

En effectuant l'approximation de la fonction $I_t^c(x,y;a)$ par un développement de Taylor du premier ordre, on obtient, en différentiant l'équation (2.9) et en tenant compte de l'expression (2.17) :

$$\nabla_a I_t^c(x,y;a) \approx M^\tau(x,y)\nabla I_t(x,y)\,. \tag{2.18}$$

En substituant les expressions (2.17) et (2.18) dans l'équation (2.16), on obtient alors :

$$\left(\sum_{(x,y)\in\Re} w_t(x,y)\,\nabla I_t^\tau(x,y)\; M(x,y)\; M^\tau(x,y)\,\nabla I_t(x,y)\right) a = -\sum_{(x,y)\in\Re} w_t(x,y)\,\frac{\partial I_t(x,y)}{\partial t}\; M^\tau(x,y)\,\nabla I_t(x,y)\,, \tag{2.19}$$

où, de manière plus compacte :

$$\left(\sum_{(x,y)\in\Re} w_t(x,y)\,\nabla_a I_t^{c\tau}(x,y)\;\; \nabla_a I_t^c(x,y;a)\right) a = -\sum_{(x,y)\in\Re} w_t(x,y)\,\frac{\partial I_t(x,y)}{\partial t}\;\; \nabla_a I_t^c(x,y;a)\,. \tag{2.20}$$

Cela permet d'exprimer le vecteur de paramètres optimal $\hat{a}$ comme solution du système linéaire (2.20) :

$$\hat{a} = \aleph^{-1}\gamma\,, \tag{2.21}$$

avec les notations suivantes :

$$\aleph = \left(\sum_{(x,y)\in\Re} w_t(x,y)\,\nabla_a I_t^{c\tau}(x,y;a)\;\; \nabla_a I_t^c(x,y;a)\right), \tag{2.22}$$

$$\gamma = -\sum_{(x,y)\in\Re} w_t(x,y)\,\frac{\partial I_t(x,y)}{\partial t}\;\; \nabla_a I_t^c(x,y;a)\,. \tag{2.23}$$

L'équation (2.23) représente en fait la solution commune à plusieurs techniques d'estimation proposées dans la littérature qui, en partant de formulations légèrement différentes, arrivent, par le biais de différentes approximations effectuées, exactement au même résultat.

La formulation la plus ancienne de l'équation (2.21) remonte au travail de pionnier de Lucas et Kanade [Lucas81].

Dans [Odobez95, Sawhney95], les auteurs effectuent d'emblée l'approximation (2.9) et proposent de minimiser directement l'expression suivante :

$$\sum_{(x,y=\in\Re}\left[\frac{\partial I_t(x,y)}{\partial t} + \nabla I_t^{\tau}(x,y)\,v(x,y;a)\right]^2. \tag{2.24}$$

Dans [Dufaux00, MPEG-4, Smolic99], l'expression (2.3) est minimisée en utilisant un algorithme de Gauss, consistant à approcher la fonctionnelle d'énergie (2.3) par une forme quadratique. Néanmoins, comme les dérivées du deuxième ordre intervenant dans le calcul de la matrice Hessienne sont négligées pour des raisons comme celles montrées dans [NRC92], la solution finale est exactement celle donnée par l'équation (2.21).

Dans [Smolic99], les auteurs construisent un système d'équations linéaire, obtenu en écrivant l'équation du flot optique pour chaque pixel de la région d'intérêt.

$$\left\{\frac{\partial I_t(x,y)}{\partial t} + \nabla I_t^{\tau}(x,y)\,v(x,y;a) = 0\right\}_{(x,y)\in\Re}. \tag{2.25}$$

Ils proposent ensuite de le résoudre par la méthode de SVD. Il est connu que la solution SVD minimise l'erreur quadratique moyenne et donc la solution aboutit à celle donnée par l'équation (2.21), dans le cas d'une fonction distance $L_2$.
Les différences entre ces approches sont donc liées principalement :

- aux différents choix des fonctions distance $\rho$ utilisées : estimateur de Tuckey dans [Odobez95], estimateur $L_2$-limité dans [Dufaux00, MPEG-4, MPEG-7_XM], estimateur de Geman McLure dans [Sawhney95],
- aux implantations mises en œuvre lors du calcul des gradients spatio-temporels, ainsi qu'à quelques détails comme l'introduction d'un facteur additif de variation de luminance $\xi$ entre deux images successives dans [Odobez95] (estimé conjointement avec les paramètres du mouvement) ou le passage dans [MPEG-7_XM] à une méthode de Levenberg-Marquardt assurant à chaque pas que l'estimé actualisé fait décroître la valeur de la fonctionnelle d'énergie.

L'approximation (2.9) nous a donc permis de re-formuler le problème de minimisation d'une fonctionnelle d'énergie fortement non-linéaire en une simple résolution d'un système d'équations linéaire. Toutefois, en pratique, et plus spécifiquement pour des mouvements de grande amplitude, l'approximation donnée par l'équation (2.9) n'est pas réaliste. Pour cette raison, on applique un algorithme incrémental d'estimation, intégré dans un schéma multirésolution de type pyramide gaussienne d'images [Burt83].
Présentons tout d'abord le principe de l'estimation incrémentale, tel que décrit dans [Lucas81, Odobez95, MPEG-4].

### 2.3.3 Estimation incrémentale

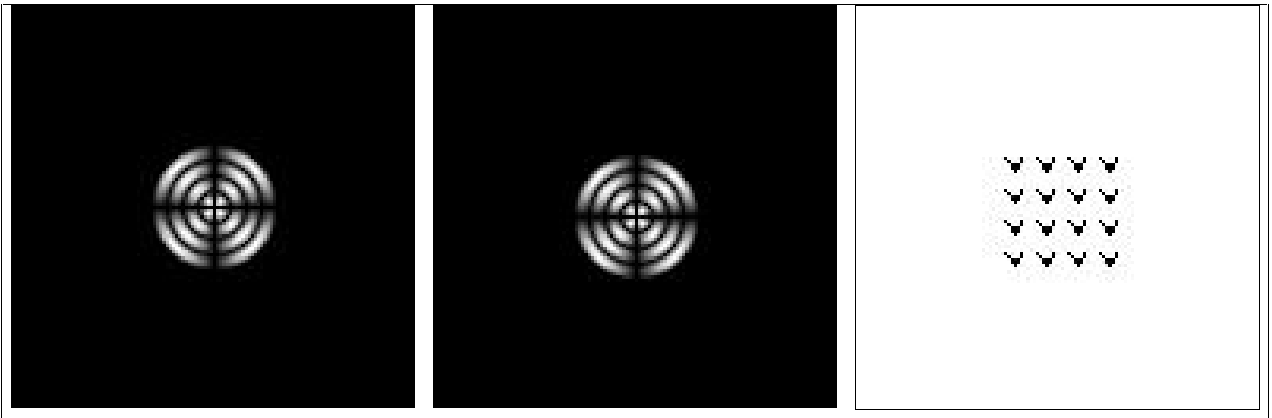
Les paramètres de mouvement donnés par la relation (2.21) sont valides uniquement si le développement au premier ordre en série de Taylor (équation (2.9)) fournit une "bonne" approximation de la fonction $I_t^c(x, y; a)$. Lors de déplacements importants, correspondant à des vecteurs vitesse supérieurs à 1 ou 2 pixels, cela n'est plus vrai.

L'idée est alors de considérer le vecteur de paramètres $\hat{a}$ estimé par (2.21) comme une première approximation de la solution, de ré-estimer par la même technique un vecteur $\delta\hat{a}$, cette fois entre l'image compensée $I_t(x, y; \hat{a})$ et $I_{t+1}(x, y)$, d'actualiser l'estimation initiale et de réitérer ce procédé, soit jusqu'à convergence (lorsque la modification du vecteur $\hat{a}$ devient négligeable), soit jusqu'à un nombre préétabli d'itérations.
Plus précisément, il s'agit d'appliquer l'algorithme suivant, exprimé ci-dessous en pseudo-code :

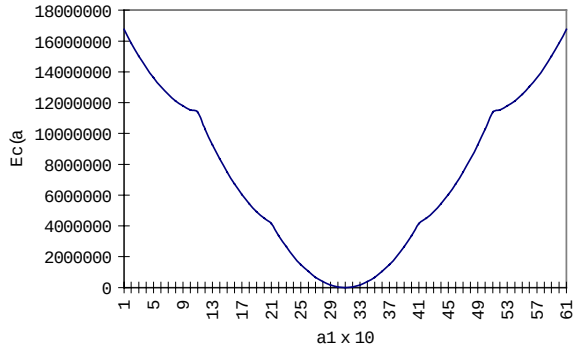
$k = 0$ ; $\hat{a} = 0$ /* Initialisation*/
**do** {
$\delta\hat{a} = \aleph^{-1}(x + v_x(x, y; \hat{a}), y + v_y(x, y; \hat{a}))\ \gamma(x + v_x(x, y; \hat{a}), y + v_y(x, y; \hat{a}))$ /* Calcul de l'incrément $\delta\hat{a}$ */
$\hat{a} = \hat{a} + \delta\hat{a}$; /* Actualisation de l'estimation */
$k$++ ;
} **while**( $k < Max_{it}$ && $\| \delta\hat{a} \|_a \geq T_a$ )

Ici, $T_a$ est un seuil fixé, $Max_{it}$ est le nombre maximal d'itérations, $\|\cdot\|_a$ désigne une norme à préciser dans l'espace des paramètres.

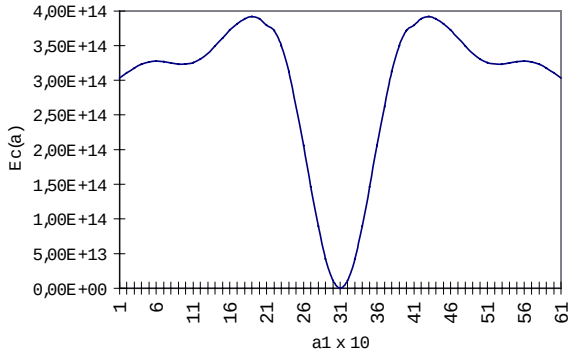
La Figure 2.5 présente l'évolution des différents termes intervenant dans le schéma d'estimation d'une translation verticale de trois pixels modélisée par un mouvement affine.



a. Deux trames successives de la séquence et le flot optique associé.

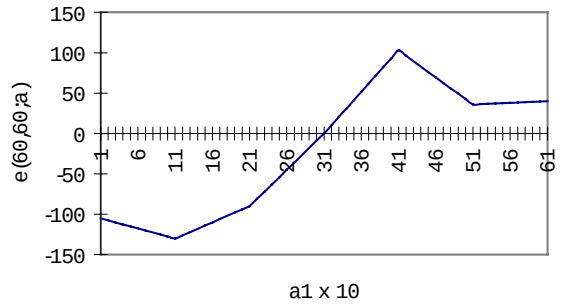



b1. Estimateur quadratique.

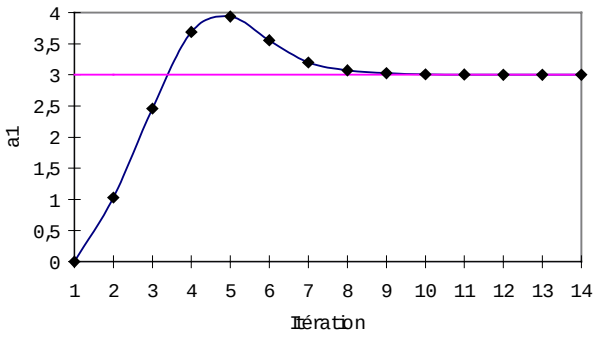



b2. Estimateur de Tuckey.

b. Comportement de la fonctionnelle d'énergie globale $E_c(a)\big|_{a_i=0,\, i\neq 1}$.




c. Comportement de $e_t(x,y;a)\big|_{\substack{(x,y)=(60,60)\\ a_i=0,\, i\neq 1}}$.




d. Evolution des valeurs estimées du paramètre $a_1$.

**Figure 2.5.** Illustration du schéma d'estimation incrémentale pour une séquence synthétique correspondant à une translation verticale de 3 pixels.

Les différentes courbes illustrées Figure 2.5.b-c montrent que les hypothèses aussi bien de convexité des fonctionnelles d'énergie que d'approximation linéaire au premier ordre de l'erreur de compensation $e(x,y;a)$ ne sont valables que dans un voisinage très restreint de la solution recherchée ($\pm 1$ pixel dans ce cas particulier de mouvement). La comparaison des fonctionnelles d'énergie quadratique et quadratique pondérée montre à l'évidence que les techniques robustes sont applicables uniquement lorsqu'une bonne initialisation est disponible, en raison de la forte non-convexité des fonctionnelles considérées. Cela augmente le risque de blocage de l'algorithme dans des minima locaux, surtout lors des premières itérations, lorsque les erreurs de compensation sont grandes et les paramètres de mouvement encore éloignés de la solution. Pour éviter cela, la largeur de la fonction de poids, contrôlée par le paramètre $C$ est ajustée à chaque itération, en utilisant par exemple le mécanisme proposé en [Odobez95]. Ainsi, au début des itérations et aux niveaux les plus grossiers de la pyramide, $C$ reçoit-il une valeur plus grande, qui va ensuite décroissant (par multiplication après chaque itération avec un facteur constant sous-unitaire) pour atteindre un niveau minimal $C_{min}$.

La Figure 2.5.d démontre la pertinence de l'algorithme incrémental d'estimation. Même si la première estimation reste très éloignée de la solution finale ($a_1 = 1.14$), les itérations successives arrivent à converger vers la solution, ici en 14 pas (avec un seuil $T_a$ de 0.0001 pour une norme $L_2$ dans l'espace des paramètres).

Tout cela, démontre objectivement l'intérêt de disposer d'une initialisation des paramètres proche de la solution optimale. Afin d'améliorer cette initialisation et de garantir la robustesse des estimations en présence de mouvements de forte amplitude, l'estimation est effectuée selon un schéma multirésolution de type pyramide gaussienne d'images [Burt83], dont le principe est rappelé ci-après.

### 2.3.4 Estimation multirésolution avec projection des paramètres

Le principe de l'estimation multi-résolution consiste à appliquer l'estimation incrémentale (*cf.* Paragraphe 2.3.3) à chaque niveau de la pyramide gaussienne et de projeter chaque fois les paramètres estimés à la résolution considérée sur le niveau de résolution immédiatement inférieur, et d'initialiser le processus d'estimation incrémentale à l'aide de cette projection.

La construction de la pyramide gaussienne d'images (Figure 2.6), assez proche dans son principe du développement des fonctions sur une base d'ondelettes, est effectuée en deux temps :

1. Filtrage passe-bas de l'image avec un noyau de convolution gaussien,
2. Sous-échantillonnage dyadique selon chaque dimension de l'image (horizontale et verticale).

a. $n = 0$. b. $n = 1$. c. $n = 2$.

**Figure 2.6.** Pyramide gaussienne à trois niveaux de résolution pour une image de la séquence MIT.

L'algorithme pyramidal d'estimation est illustré Figure 2.7. Il est initialisé au niveau le plus grossier de la pyramide. Une fois la convergence atteinte au niveau *n* courant, les paramètres sont projetés et utilisés comme initialisation de l'algorithme d'estimation incrémentale au niveau inférieur (*n-1*), selon la règle suivante :

$$P_a : a_i^{ini,n-1} = 2a_i^{fin,n}, \quad si \;\; i \in \{1,2\} \qquad \text{et} \qquad a_i^{ini,n-1} = a_i^{fin,n}, \quad si \;\; i \notin \{1,2\}, \tag{2.26}$$

où

- $P_a$ désigne l'opérateur de projection des paramètres,
- $\left\{a_i^{fin,n}\right\}_i$ sont les paramètres finaux estimés à la résolution *n*,
- $\left\{a_i^{ini,n-1}\right\}_i$ sont les paramètres initiaux à la résolution *n*-1.

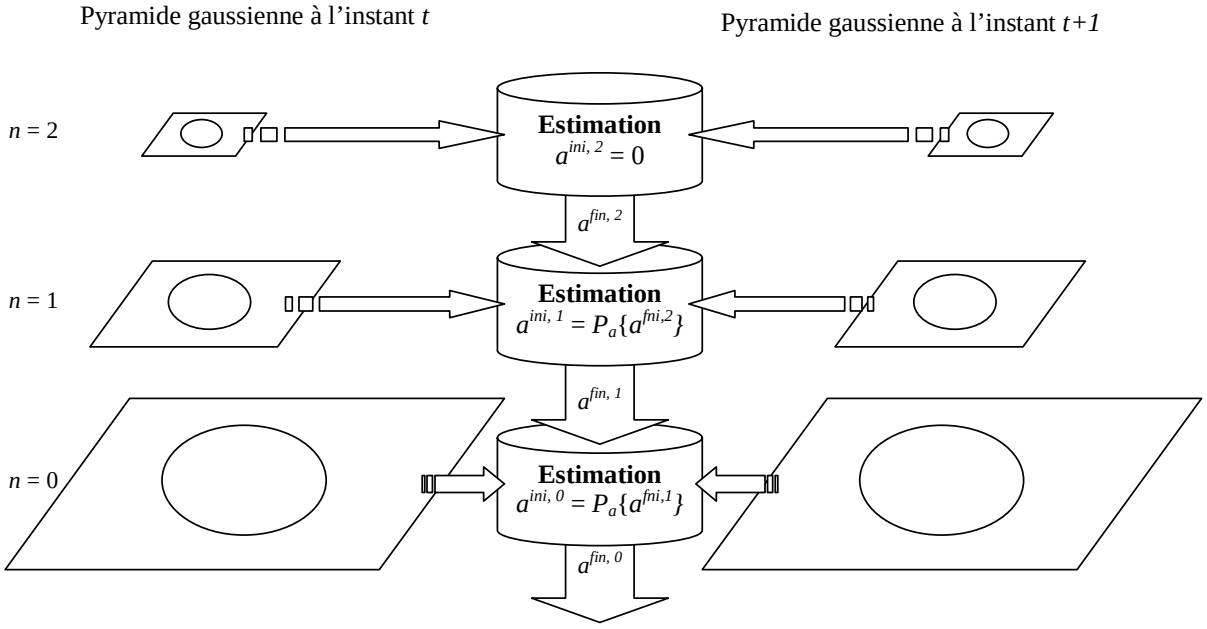
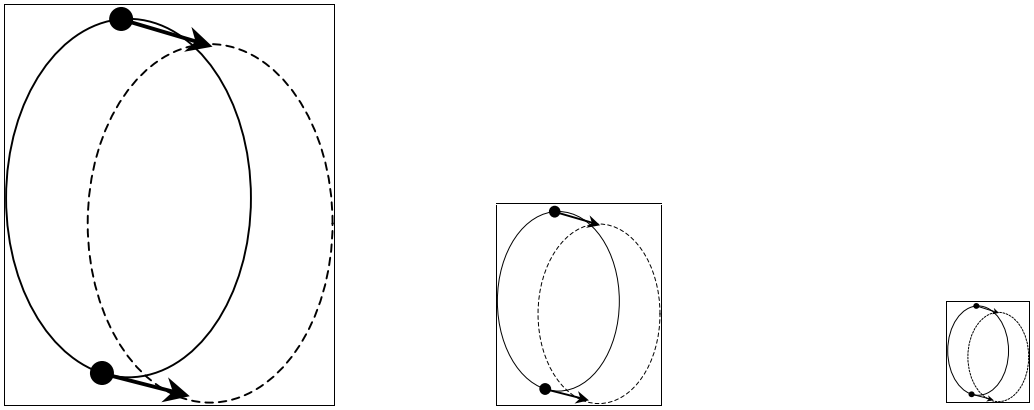



**Figure 2.7.** Schéma de principe de l'algorithme d'estimation multirésolution.

Ce schéma permet de conditionner correctement l'équation du flot optique (2.9) et de gérer des déplacements de grande amplitude (Figure 2.8). De plus, la projection des paramètres d'un niveau à l'autre permet à chaque étape d'obtenir une initialisation pas trop éloignée de la solution recherchée.

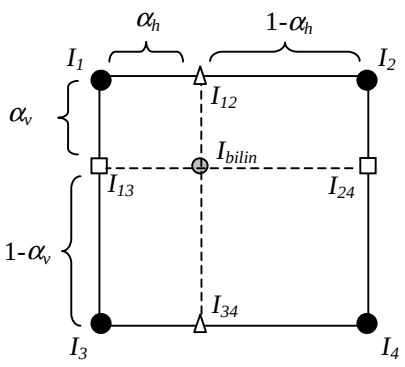


a. $n = 0$. b. $n = 1$. c. $n = 2$.

**Figure 2.8.** Décroissance exponentielle des amplitudes des vecteurs vitesse en fonction du niveau croissant de hiérarchie.

Une étape de l'estimation des paramètres concerne le calcul des gradients spatio-temporels. En effet, comme le calcul des dérivées est très sensible au bruit, nous avons adopté la méthode d'estimation des gradients de Canny [Canny86], qui effectue un filtrage de l'image par un noyau de convolution gaussien, et qui estime ensuite les dérivées par différences finies.

Remarquons également que tous les développements effectués aux Paragraphes 2.3.2 et 2.3.3 s'inscrivent dans un cadre mathématique continu. Lorsqu'il s'agit d'images discrètes, les valeurs des fonctions $I_t(x,y)$ sont uniquement connues pour des valeurs entières de *x* et de *y*. Pour calculer la fonction $I_t^c(x,y;a)$ pour tout vecteur de paramètres *a*, il est alors nécessaire d'interpoler la valeur de la luminance $I_t$ au point $(x+v_x(x,y;a), y+v_y(x,y;a))$ à partir des valeurs entières les plus proches. Pour des raisons de simplicité, nous avons choisi ici d'utiliser une interpolation bi-linéaire, illustrée Figure 2.9.



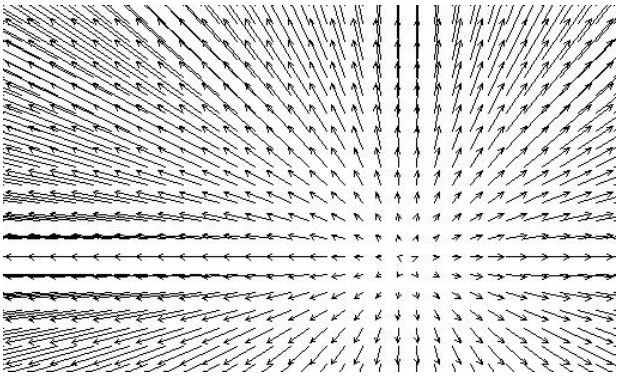

$$I_{12} = (1-\alpha_h)I_1 + \alpha_h I_2$$
$$I_{34} = (1-\alpha_h)I_3 + \alpha_h I_4$$
$$I_{13} = (1-\alpha_v)I_1 + \alpha_v I_3$$
$$I_{24} = (1-\alpha_v)I_2 + \alpha_v I_4$$
$$I_{bilin} = (1-\alpha_h)I_{13} + \alpha_h I_{24} = (1-\alpha_v)I_{12} + \alpha_v I_{34}$$

**Figure 2.9.** Principe de l'interpolation bilinéaire.

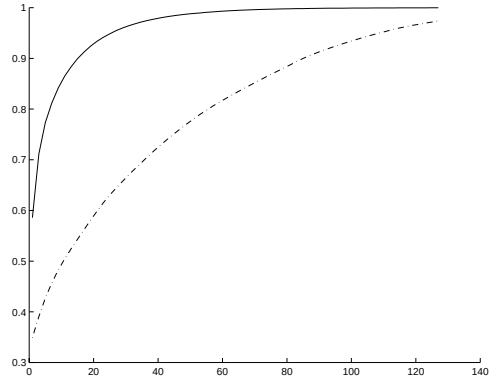
### 2.3.5 Exemples d'estimation

Afin d'illustrer la grande variété de types de mouvement pris en compte par les modèles paramétriques et les performances de l'algorithme d'estimation implanté, présentons quelques exemples, correspondant à des mouvements aussi bien globaux que locaux.
Commençons avec la séquence bien connue du MIT (Figure 2.10).

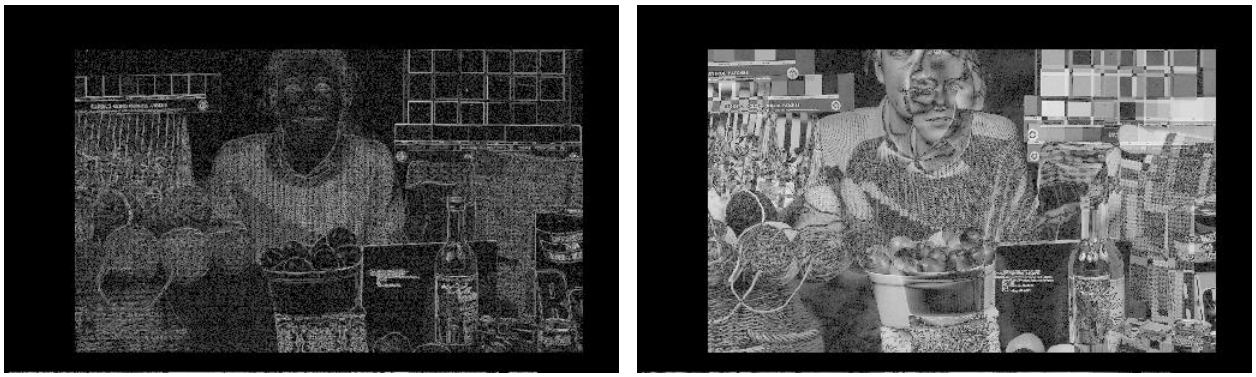
a. Deux trames successives de la séquence MIT



b. Flot optique associé au modèle affine estimé.




c. Histogramme cumulé des valeurs absolues des différences d'images sans (en pointillés) et avec (en trait plein) compensation affine de mouvement.



d. Différences des deux trames a et b avec (à gauche) et sans (à droite) compensation affine de mouvement.

**Figure 2.10.** Estimation du mouvement de la caméra.

Il s'agit là d'un mouvement global correspondant à un zoom-avant qui est correctement estimé, malgré son amplitude importante, comme le montre la Figure 2.10.d, qui visualise les différences des deux

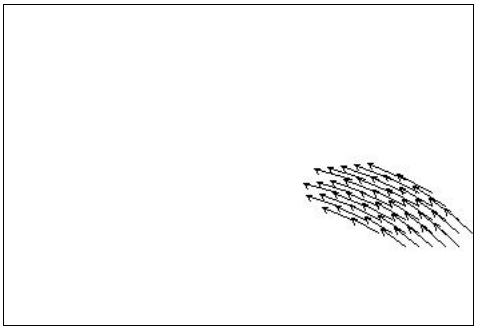
trames (sur la région de recouvrement), avec et sans compensation de mouvement (avec une correction gamma d'un facteur 3.0 pour une meilleure visualisation).

Poursuivons avec la catégorie spécifique de mouvements purement 2D. Il s'agit d'estimer entre les trames successives différents mouvements d'objets individuels, les masques binaires de ceux-ci résultant d'un procédé de segmentation spatiale ou spatio-temporelle. Les Figures (Figure 2.11-Figure 2.15) présentent les images entre lesquelles l'estimation est effectuée (a), le flot optique associé au modèle estimé (b) et l'histogramme cumulé des différences absolues des images avec (en trait plein) et sans (en pointillés) compensation de mouvement (c).

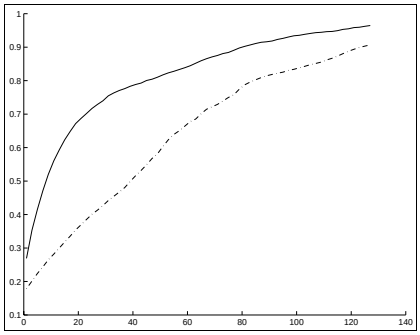
La Figure 2.11 illustre un mouvement de rotation avec un centre de rotation très éloigné de l'objet (approximativement 3 fois la taille de l'objet) correspondant à une voiture en virage. Ici, l'objet a été segmenté manuellement sur la première trame de la séquence.

Les mouvements suivants correspondent à différents gestes de la main droite sélectionnés à partir de la base de mouvements naturels INT (*cf.* Paragraphe 2.5.5). La Figure 2.12 présente une rotation (approximativement 20°) dont le centre est inclus dans la région support de l'objet, comme le montre bien l'image de flot optique. Ici aussi, malgré sa grande amplitude, le mouvement est correctement estimé. Les Figures (Figure 2.13 et Figure 2.14) présentent respectivement des exemples de zoom-arrière et zoom-avant. Enfin, la Figure 2.15 illustre l'exemple d'un mouvement translationnel pur. L'amplitude du vecteur de translation estimé est de 72 pixels, ce qui est en parfaite cohérence avec le mouvement réel de la main dans la vidéo. Cela démontre bien l'efficacité du schéma d'estimation multi-résolution adopté.

a. Deux trames successives.



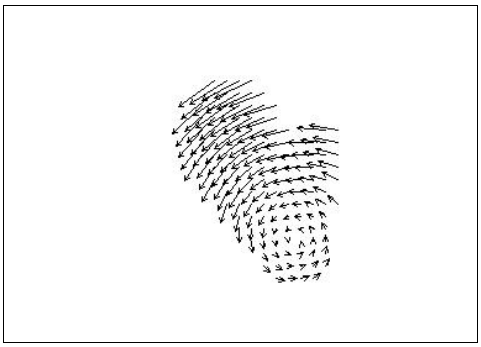
b. Flot optique associé au modèle affine estimé.



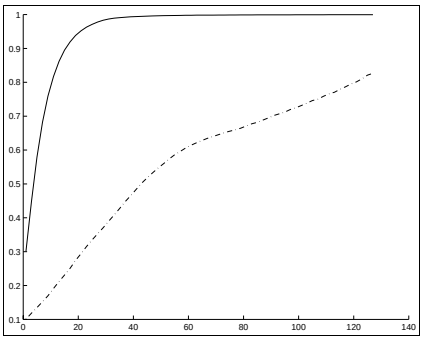

c. Histogramme cumulé des valeurs absolue des différences d'images sans (en pointillés) et avec (en trait plein) compensation affine de mouvement.

**Figure 2.11.** Estimation par mouvement affine pour une rotation de centre éloigné.

a. Deux trames successives.



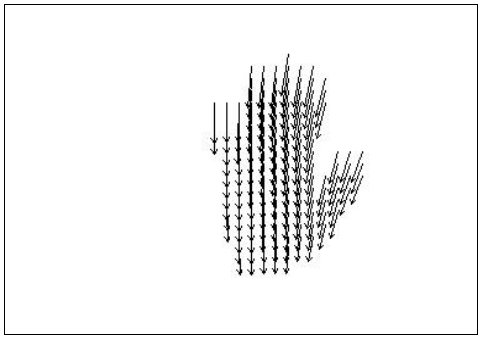
b. Flot optique associé au modèle affine estimé.



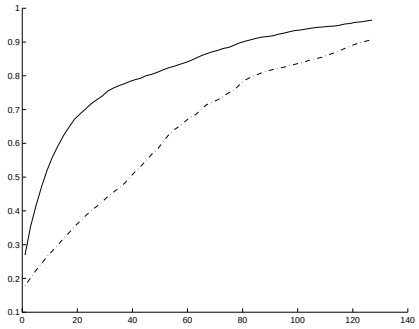

c. Histogramme cumulé des valeurs absolue des différences d'images sans (en pointillés) et avec (en trait plein) compensation affine de mouvement.

**Figure 2.12.** Estimation par mouvement affine d'une rotation de la main.

a. Deux trames successives.

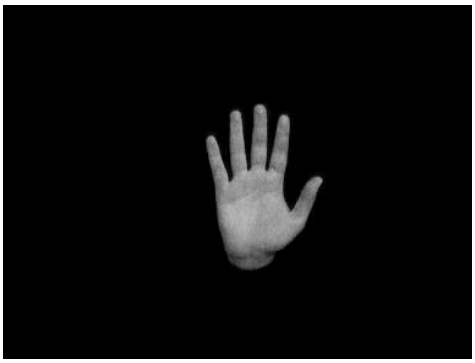


b. Flot optique associé au modèle affine estimé.

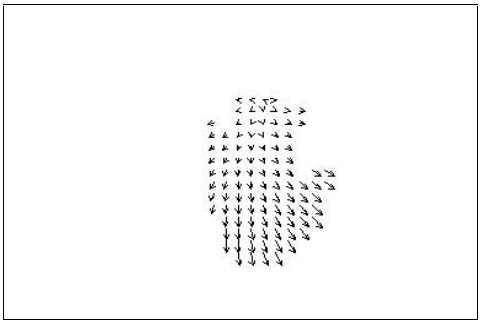



c. Histogramme cumulé des valeurs absolue des différences d'images sans (en pointillés) et avec (en trait plein) compensation affine de mouvement.

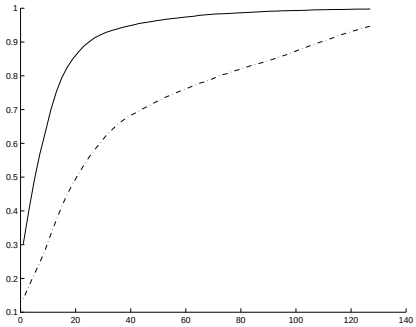
**Figure 2.13.** Estimation par mouvement affine d'un zoom-arrière combiné à une translation vers le bas.




a. Deux trames successives.

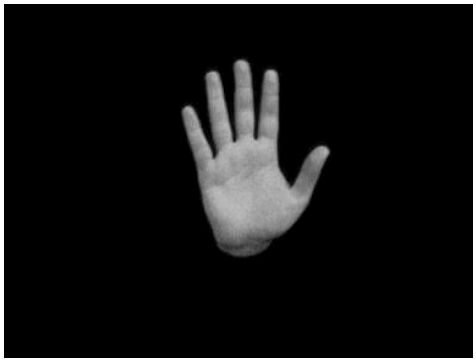
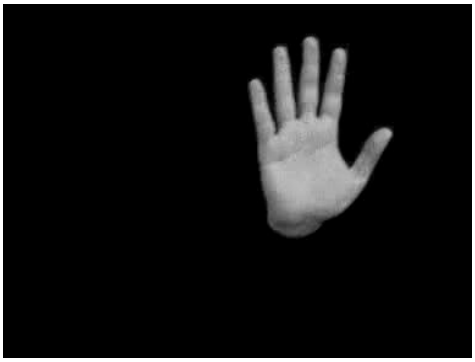


b. Flot optique associé au modèle affine estimé.

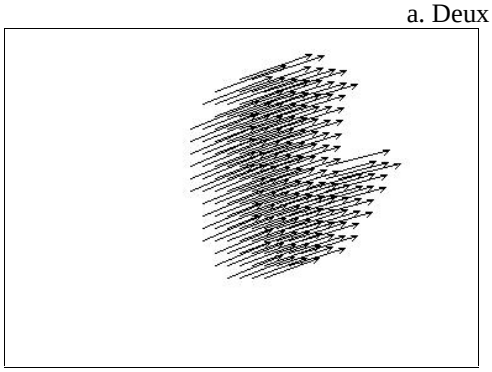



c. Histogramme cumulé des valeurs absolue des différences d'images sans (en pointillés) et avec (en trait plein) compensation affine de mouvement.

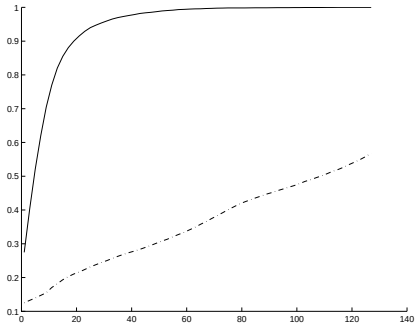
**Figure 2.14.** Estimation de mouvement affine pour un mouvement de zoom-avant.

a. Deux trames successives.



b. Flot optique associé au modèle affine estimé.



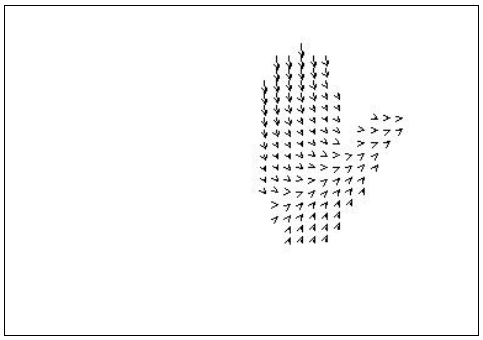

c. Histogramme cumulé des valeurs absolue des différences d'images sans (en pointillés) et avec (en trait plein) compensation affine de mouvement.

**Figure 2.15.** Estimation par mouvement affine d'une translation de grande amplitude (72 pixels).

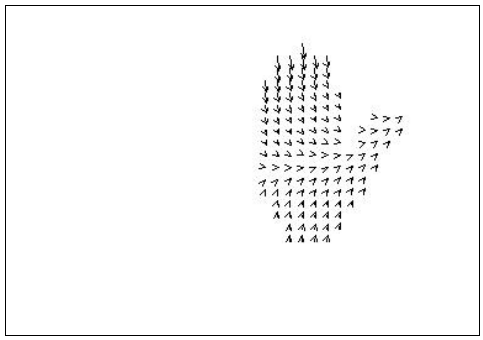
Pour tous ces mouvements 2D purs, la modélisation affine est naturellement bien adaptée. Toutefois, dans le cas de mouvements plus complexes, comme ceux provenant de la projection en 2D des mouvements d'objets évoluant dans une scène 3D, il est intéressant d'établir si ce bon comportement du modèle affine persiste.

La Figure 2.16 présente l'exemple d'un mouvement 3D horizontal, avec les champs de vitesse et les cartes de différences compensées correspondant à un modèle affine et à un modèle quadratique. Les deux images de flot optique montrent clairement la supériorité du modèle quadratique par rapport au modèle affine. Le flot optique quadratique présente une ligne horizontale, passant par le milieu de l'objet et ayant des vitesses nulles, ce qui est caractéristique de ce type de mouvement. Dans le cas du flot optique affine, cette ligne est légèrement déviée. Les images de différences compensées prouvent encore plus cette supériorité, en faisant ressortir plus particulièrement les erreurs obtenues sur les bords de l'objet. Ces commentaires restent également valables pour le mouvement 3D vertical présenté Figure 2.17.

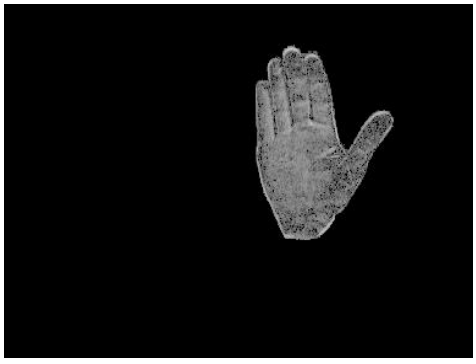
a. Deux trames successives.



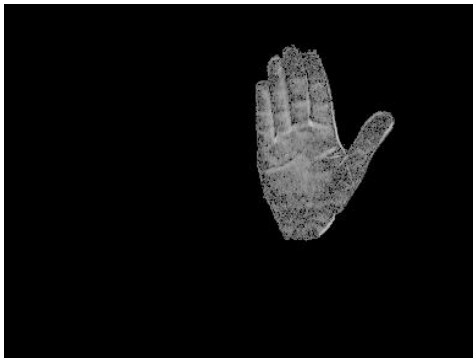
b. Flot optique associé au modèle affine estimé



c. Flot optique associé au modèle quadratique estimé.

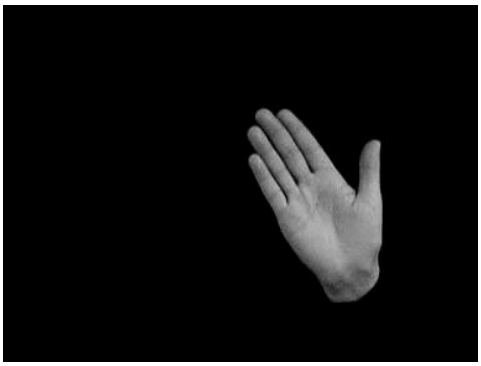


d. Images de différences compensées pour un modèle affine.

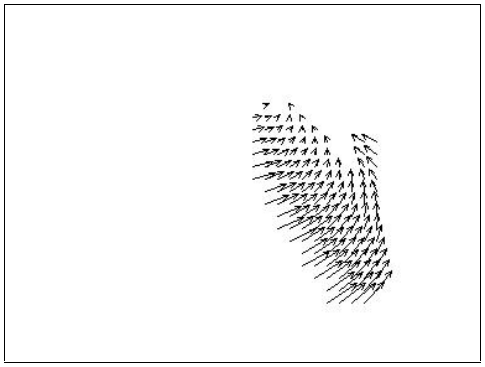


e. Images de différences compensées pour un modèle quadratique

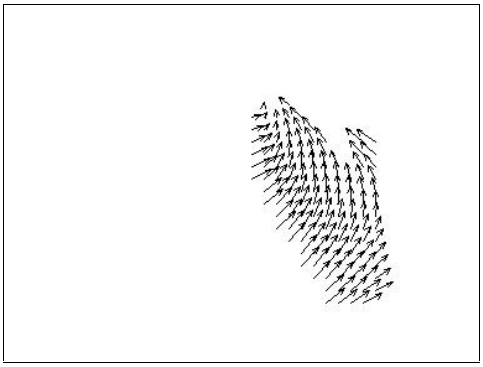
**Figure 2.16.** Mouvement 3D horizontal.



a. Deux trames successives.



b. Flot optique associé au modèle affine estimé



c. Flot optique associé au modèle quadratique estimé.

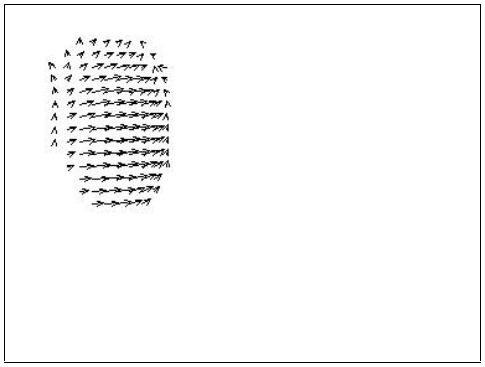
d. Images de différences compensées pour un modèle affine.

e. Images de différences compensées pour un modèle quadratique

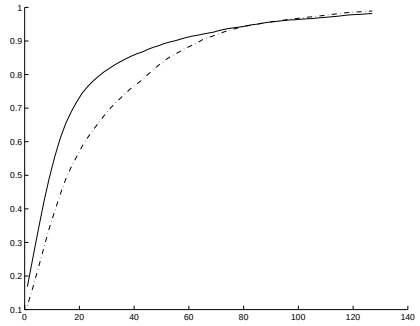
**Figure 2.17.** Mouvement 3D vertical.

Un dernier exemple, cette fois pour un mouvement vertical de tête et avec une modélisation perspective planaire est présenté Figure 2.18. Ici, le modèle affine n'a même pas réussi à atteindre la convergence. Remarquons la fidélité du mouvement estimé, malgré la grande amplitude du mouvement et la présence des régions d'apparition et de recouvrement sur les bords de l'objet.

a. Deux trames successives.



b. Flot optique associé au modèle estimé.




c. Histogramme cumulé des valeurs absolues des différences d'images sans (en pointillés) et avec (en trait plein) compensation affine de mouvement.

**Figure 2.18.** Estimation d'un mouvement de tête par un modèle perspectif planaire.

Tous ces exemples montrent que la modélisation affine est surpassée par les modèles d'ordre supérieur en terme de fidélité du mouvement estimé par rapport au mouvement réel des objets. Néanmoins, elle offre des approximations relativement "raisonnables" des mouvements 2D présents dans les images et représente donc un bon compromis entre fidélité de la représentation et complexité de calcul, qui augmente avec la complexité du modèle considéré.
Remarquons également que malgré la nature 3D de ces différents mouvements, les modèles estimés approchent avec fidélité les champs de vitesse apparente. Cela est dû au fait que les objets de la scène sont suffisamment éloignés de la caméra et satisfont donc une certaine hypothèse de planéité, au sens où les discontinuités de profondeur peuvent être négligées. Bien évidemment, la modélisation par modèle paramétrique 2D cesse d'être efficace dès que ces conditions ne sont plus respectées, mais ce cas ne sera pas traité dans ce travail, où nous nous intéressons uniquement à la classe suffisamment large de mouvements pouvant être décrits avec fidélité par modèles paramétriques 2D.

Enfin, afin d'illustrer l'influence des approches par estimateur robuste, nous avons tout d'abord considéré les mêmes séquences de geste, en sélectionnant cette fois des mouvements présentant des composantes non-affines plus ou moins importantes. Dans ces exemples, le modèle considéré est toujours affine.

La Figure 2.19 présente une déformation d'amplitude relativement faible. L'estimation non-robuste moyenne, en un certain sens, l'ensemble des mouvements individuels des pixels de l'objet et conduit à un flot optique n'ayant que peu de rapport avec les mouvements réels. Cela est dû en partie à la relative

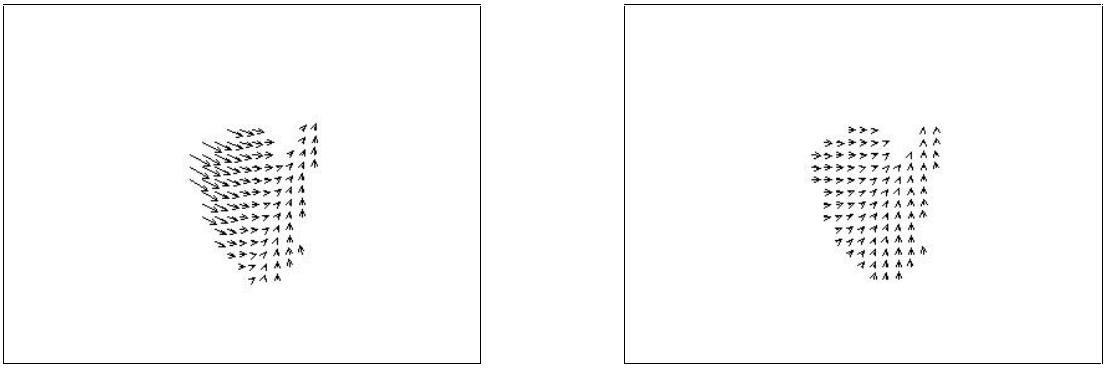
uniformité des niveaux de gris des mains : le critère de minimisation étant fondé sur l'erreur de compensation, des pixels dont les vecteurs vitesse sont mal estimées peuvent néanmoins conduire à des erreurs de compensation petites. En revanche, en appliquant un estimateur robuste, les mouvements sont correctement estimés, au moins pour l'ensemble des pixels de la région de la paume, pour laquelle les déformations sont quasi nulles. Ce phénomène est nettement visible sur les images des différences compensées.
Les estimateurs robustes prennent également en compte les changements de luminance dus aux effets d'ombrage, typiques pour ce type de mouvement non-affine, car les poids associés à ces régions sont nuls (Figure 2.19.d).

La Figure 2.20 montre un deuxième exemple, présentant une déformation très importante : quatre doigts recouvrent une large partie de la paume et le pouce a un mouvement indépendant de ceux des doigts. A nouveau, le mouvement des région rigides est mieux pris en compte par l'estimateur robuste.

Des déformations combinées à la présence de recouvrements et d'apparitions, typiques des mouvements 3D, sont illustrées dans l'exemple de la Figure 2.21. Ici, le pouce présente un mouvement indépendant du reste de la main, une partie du poignet apparaît alors qu'une partie de la paume est occultée par les doigts. Les flots optiques associés aux techniques robustes et non-robustes sont visuellement bien similaires, ce qui prouve que l'estimation est raisonnable dans les deux cas. Les images des différences compensées démontrent néanmoins la supériorité de l'estimateur robuste. Les erreurs de compensation sont bien plus petites avec un estimateur robuste qu'avec un estimateur non-robuste dès lors qu'il s'agit des pixels appartenant aux régions rigides. De plus, les régions de recouvrement/apparition sont clairement visibles sur l'images des différences compensées. D'ailleurs, l'image de poids présentée Figure 2.21.d montre bien la région de l'objet sur laquelle l'estimation a été effectuée et qui correspond notamment à ces régions rigides.

Enfin, la Figure 2.22 présente un dernier exemple, correspondant cette fois à l'estimation du mouvement global de la caméra en présence d'autres objets en mouvement dans la scène. Les images de flot optique et les cartes de différences compensées montrent clairement la supériorité de l'estimateur robuste. En outre, la carte de poids fournit une localisation approximative des objets en mouvement indépendant. Notons que les vecteurs de vitesse associés au modèle affine estimé sont orientés vers la droite, contrairement donc au mouvement de l'objet (le sauteur à la perche) qui se déplace vers la gauche.

a. Deux trames successives.

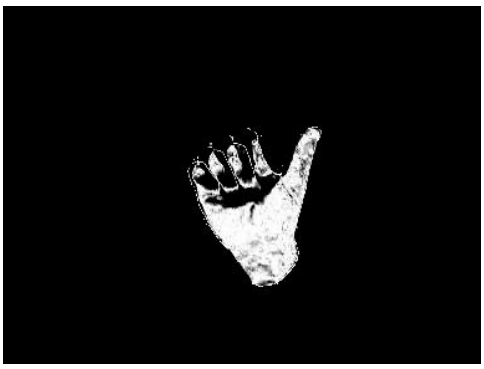


b1. Estimateur non-robuste. b1. Estimateur robuste.

b. Flot optique associé au modèle affine estimé.

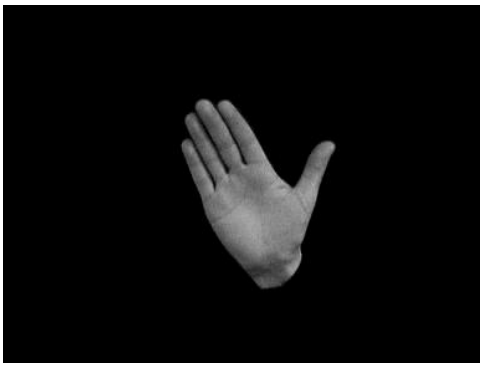
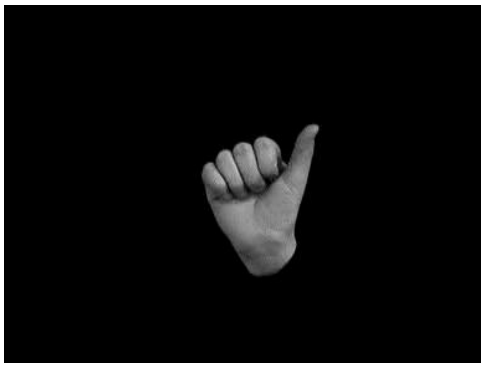
c1. Estimateur non-robuste. c2. Estimateur robuste.

c. Visualisation des différences compensées.

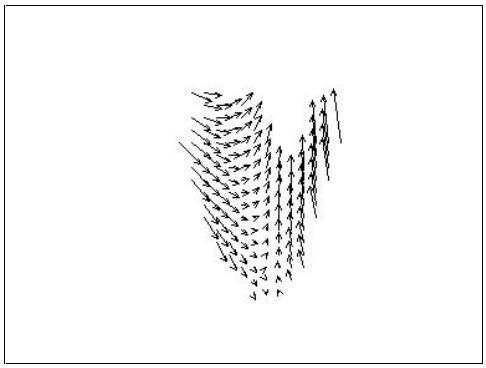
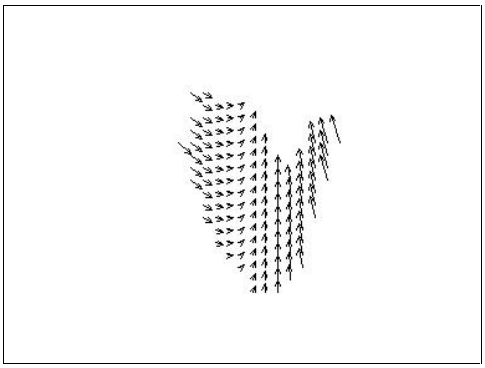


d. Visualisation des poids de l'estimateur robuste.

**Figure 2.19.** Estimation robuste et non-robuste pour une déformation relativement faible de la main.

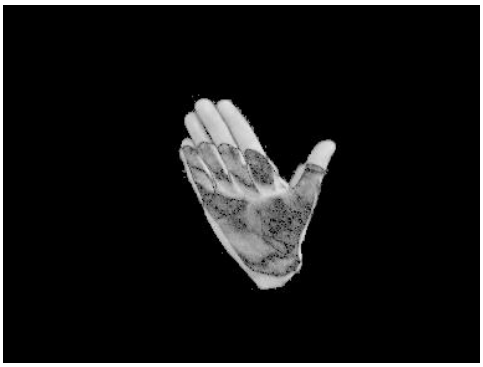
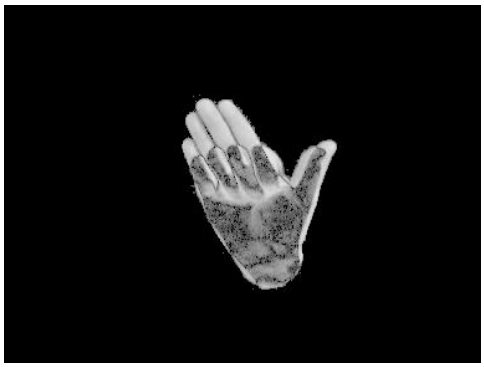



a. Deux trames successives.

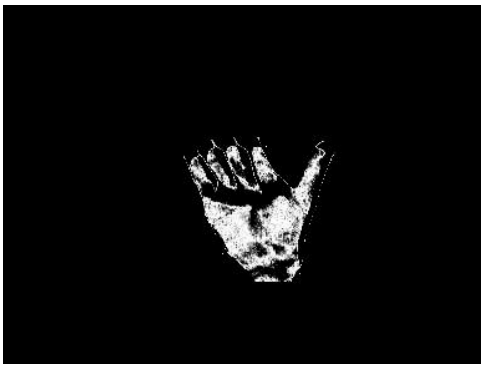
b1. Estimateur non-robuste. b1. Estimateur robuste.

b. Flot optique associé au modèle affine estimé.

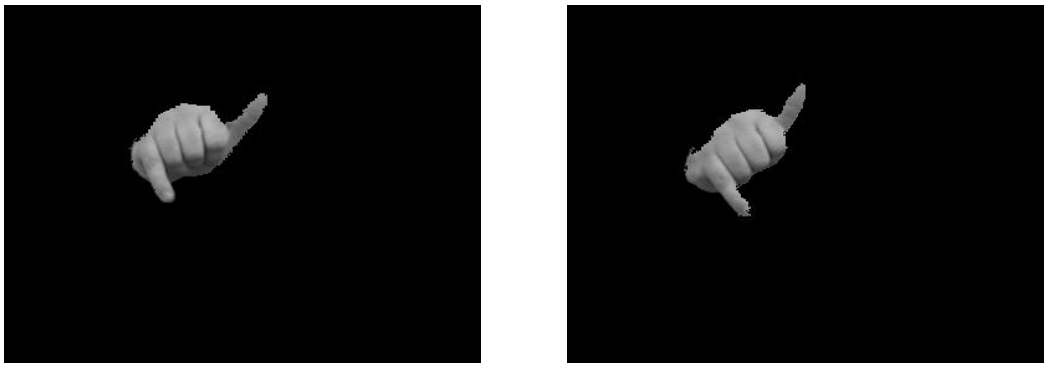
c1. Estimateur non-robuste. c2. Estimateur robuste.

c. Visualisation des différences compensées.

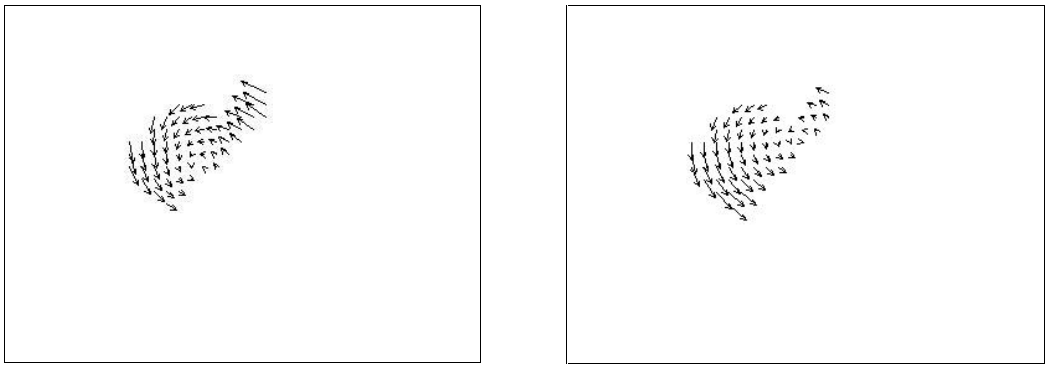


d. Visualisation des poids de l'estimateur robuste.

**Figure 2.20.** Estimation robuste et non-robuste pour une très forte déformation de la main.

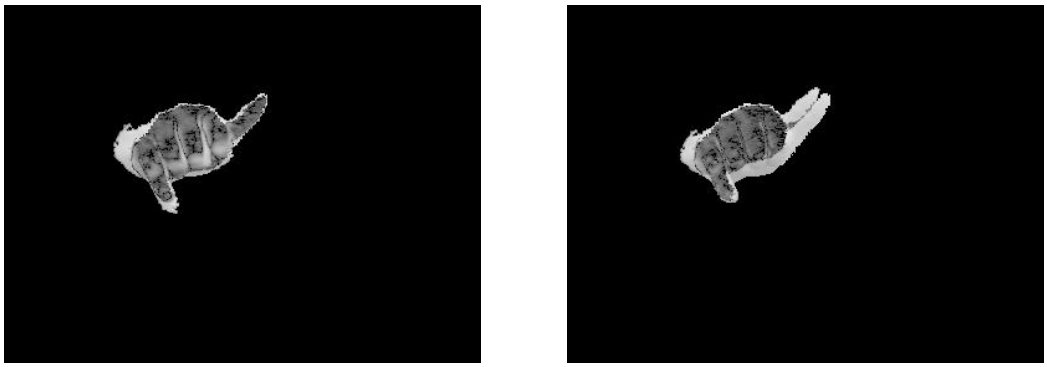


a. Deux trames successives.



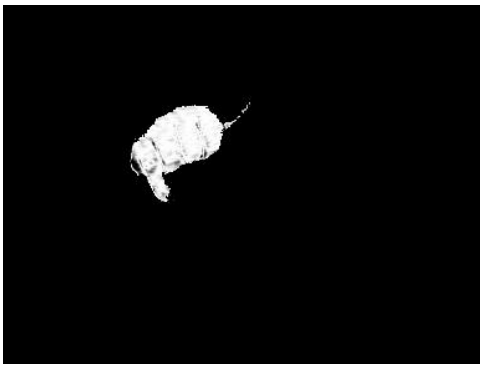
b1. Estimateur non-robuste. b1. Estimateur robuste.

b. Flot optique associé au modèle affine estimé.



c1. Estimateur non-robuste. c2. Estimateur robuste.

c. Visualisation des différences compensées.



d. Visualisation des poids de l'estimateur robuste.

**Figure 2.21.** Estimation robuste et non-robuste pour un mouvement 3D.

En conclusion, l'ensemble des résultats expérimentaux présentés permet de conclure à la pertinence de l'approche adoptée

## 2.4 Segmentation temporelle en régions de mouvement cohérent

Une région spatio-temporelle sera appelée *région de mouvement cohérent* (RMC) si on peut lui associer un unique vecteur de paramètres décrivant de manière fidèle le mouvement sur l'ensemble du support spatio-temporel considéré.

Une fois les paramètres de mouvement déterminés pour chaque paire de trames successives de la séquence, la stratégie suivante a été adoptée :

1. Calculer les mesures de similarité à base de champs de vitesses (MSSCV) (*cf.* Paragraphe 2.5.3) entre chaque paire de trames et effectuer un seuillage de ces distances pour déterminer une première sur-segmentation de la séquence,
2. Fusionner les deux RMC successives qui peuvent être décrites de manière "fidèle" par un unique modèle de mouvement. Cette deuxième étape a pour but d'éliminer l'influence de certaines aberrations susceptibles d'apparaître lorsque seuls les paramètres correspondant à des transitions successives sont considérés, alors qu'ils peuvent être altérés par du bruit ou par les conditions d'acquisition.

Notons que l'étape de fusion doit être réalisée conditionnellement à une segmentation en plans de la vidéo (supposée disponible), afin d'éviter de fusionner des RMC ayant le même mouvement mais situées dans des plans différents.

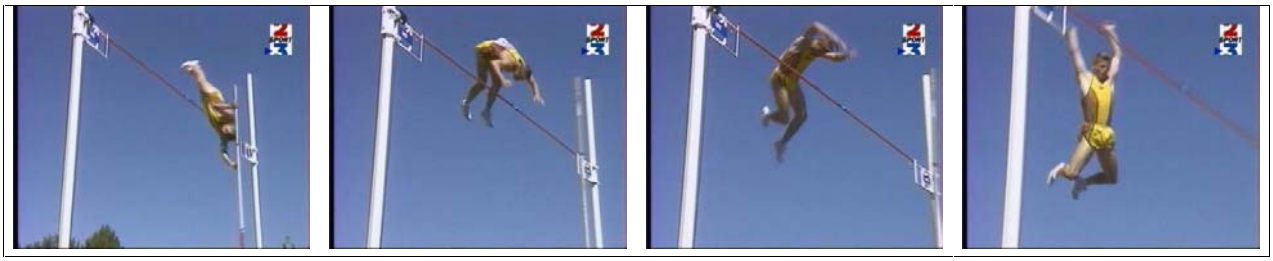
Dans ce cadre générique, plusieurs stratégies ont été considérées, correspondant à différentes manières de définir le critère de fidélité au mouvement. La façon la plus simple consiste à moyenner les paramètres de mouvement sur l'ensemble des trames correspondant à la première segmentation pour chaque intervalle temporel déterminé. Ensuite, les MSSCV sont calculées pour deux intervalles successifs. Chaque paire d'intervalles donnant une distance inférieure à un seuil fixé est ensuite fusionnée en un seul intervalle et les paramètres de mouvement caractérisant le nouvel intervalle sont recalculés comme la moyenne des paramètres sur chacun des intervalles initiaux.

Cette stratégie simple mais efficace s'est révélée particulièrement utile pour la segmentation des plans de la vidéo en RMC associées au mouvement global de la caméra. Dans la littérature, on considère habituellement le plan comme étant l'atome élémentaire de description d'une vidéo. Toutefois, du point de vue du mouvement, des RMC bien différentes peuvent se trouver à l'intérieur d'un même plan. La Figure 2.23 illustre les cinq intervalles de mouvement cohérent associés à un plan d'une séquence vidéo correspondant à une retransmission sportive de saut à la perche.
Notons en outre que pour ce type de retransmission, le mouvement global de la caméra est très fortement corrélé au mouvement des différents acteurs intervenant dans la scène, l'objectif de l'opérateur étant alors de suivre au mieux les différents sportifs. L'information de mouvement associée aux RMC est donc d'autant plus pertinente par rapport aux événements de la scène.

a. Le début de l'élan, correspondant à un déplacement vers la droite de la caméra.

b. La fin de l'élan, correspondant toujours à un déplacement vers la droite, mais avec une amplitude augmentée

c. Le saut lui-même, correspondant à un déplacement vertical ascendant



d. Le moment où le sportif franchit la barre, correspondant à un zoom-avant.

e. La chute, correspondant à un déplacement vertical descendant.

**Figure 2.23.** Les cinq intervalles de mouvement cohérent (un par ligne) associés à un plan de saut à la perche (vidéo du corpus AGIR).

Cet exemple montre bien qu'il n'existe pas une véritable homogénéité du mouvement à l'intérieur du même plan vidéo et que les RMC offre le bon cadre pour définir les éléments atomiques de mouvement

paramétrique homogène. Définir des mesures de similarité de mouvement entre deux plans doit alors tenir compte de cette décomposition de ceux-ci en RMC, en utilisant par exemple des fonctions de coût comme celles typiques à l'appariement de chaînes de caractères (*string matching*). Ces aspects de plus haut niveau dépassent largement le cadre de ce travail et ne seront pas abordés par la suite. Ici, nous intéressons uniquement à définir une mesure de similarité de mouvement entre deux RMC, chacune décrite par un unique modèle de mouvement correspondant à une succession de trames considérée comme représentative pour toute la région spatio-temporelle.

## 2.5 Mesures de similarité

Si le DMPO présente l'avantage de pouvoir représenter de manière très compacte une large classe de mouvements, son application à des requêtes par similarité de mouvement nécessite de disposer de mesures de similarité efficaces. Or, la littérature fait état d'un manque à ce sujet [Deng98].

### 2.5.1 Distances dans l'espace des paramètres

Les mesures de similarité définies dans l'espace des paramètres (MSEP) présentent les inconvénients majeurs suivants :

- La contribution spécifique de chaque paramètre n'est pas prise en compte. En effet, dans le cas des mouvements présents dans les séquences naturelles, les composantes de translation sont en général prédominantes de plusieurs ordres de grandeur par rapport aux autres composantes de mouvement. Les mesures de similarité sont alors fortement biaisées par ces composantes translationnelles. Une solution consisterait à introduire des pondérations, mais le choix des poids pour chaque paramètre relève d'heuristiques.
- De telles distances sont difficilement interprétables, puisque elles ne reposent sur aucune interprétation physique.
- Ces distances ne permettent pas la comparaison de mouvements spécifiés par différents modèles ou la comparaison d'un modèle de mouvement avec un mouvement spécifié par un champ dense de vitesse.

Afin de pallier ces inconvénients, nous proposons de construire des mesures de similarité fondées sur des fonctions distance appliquées directement sur les champs de vecteurs associés aux modèles de mouvement.

### 2.5.2 Distances entre champs de vitesse

Des distances entre champs de vitesse associés aux modèles de mouvement paramétriques ont déjà été considérées dans le contexte de la segmentation markovienne par le mouvement [Gelgon96] de manière à définir des fonctions de potentiel favorisant la fusion de régions adjacentes de mouvements cohérents.
Les vecteurs vitesse ont été également utilisés dans le cadre de la segmentation au sens du mouvement dans [Szeliski96, Zhou00]. Toutefois, les objectifs sont ici légèrement différents, puisqu'il ne s'agit pas de déterminer une mesure globale de similarité entre deux champs de vecteurs vitesse définis sur des régions support arbitraires et distinctes. En effet, dans [Zhou-JY00] les auteurs estiment deux champs de

vitesse, l'un correspondant au mouvement global dominant et l'autre extrait à l'aide d'un algorithme de flot optique dense. Ils proposent ensuite de calculer les différences entre les vecteurs vitesse des deux champs en chaque pixel et de les seuiller pour déterminer les régions supports des objets ayant un mouvement différent de celui de la caméra. Dans [Szeliski96], les auteurs modélisent le mouvement dans la scène par des modèles paramétriques à base de fonctions B-splines, définies sur des blocs rectangulaires. Leur taille est adaptativement définie par subdivisions successives conditionnellement à un seuil sur le mouvement résiduel [Irani92]. La norme $L_p$ du champ de vitesse sur l'ensemble des pixels du bloc considéré est utilisée comme mesure du mouvement résiduel. La question de déterminer une mesure de similarité entre deux mouvement arbitraires définis sur des supports spatiaux différents n'est pas abordé.

Ici, notre objectif est de construire des mesures de similarité permettant la comparaison de mouvements d'objets arbitraires. Les mesures de similarité fondées sur les champs de vitesses (MSCV) [Prêteux-Iso99.06m, Zaharia-Iso99c, Zaharia-Iso99d, Zaharia01] s'appuient sur l'hypothèse qu'à des mouvements visuellement similaires doivent correspondre des champs de vitesses analogues.
Considérons deux modèles de mouvement, notés $\mathbf{M_1}$ et $\mathbf{M_2}$, associés respectivement à deux régions spatiales, $\mathbf{R_1}$ et $\mathbf{R_2}$. Tout d'abord, nous alignons les objets de telle sorte que leurs centres de gravité, $g_1$ et $g_2$, coïncident avec l'origine du repère cartésien considéré. Soient $\mathbf{R'_1}$ et $\mathbf{R'_2}$ les régions correspondantes après alignement. Nous définissons la MSCV entre les mouvements $\mathbf{M_1}$ et $\mathbf{M_2}$ comme la somme des distances entre les vecteurs vitesse correspondant sur la réunion des supports spatiaux $\mathbf{R'_1}$ et $\mathbf{R'_2}$ :

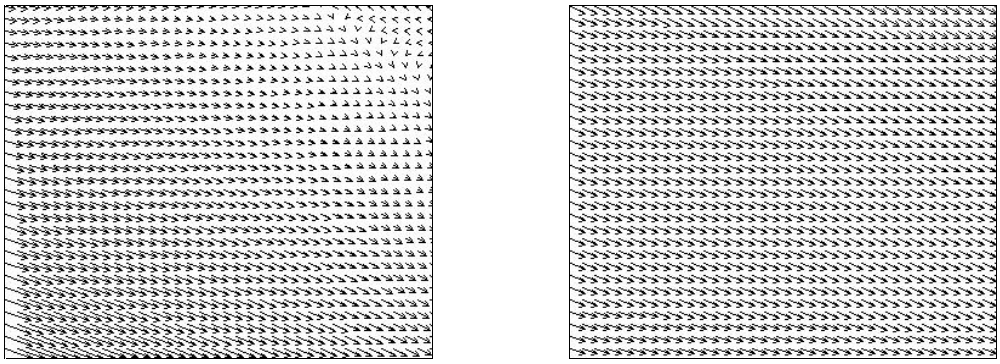
$$D(\mathrm{M}_1,\ \mathrm{M}_2) = \sum_{p \in \mathrm{R'}_1 \cup \mathrm{R'}_2} d(v_1(p+g_1;\mathrm{M}_1), v_2(p+g_2;\mathrm{M}_2)), \tag{2.27}$$

où

- $v_i(p+g_i;\mathrm{M_i})$ désigne le vecteur vitesse 2D au point $p+g_i$ correspondant au modèle de mouvement $\mathbf{M_i}$, $i \in \{1,2\}$,
- $d(v_1(p), v_2(p))$ est une fonction distance à préciser entre les deux vecteurs de vitesse.

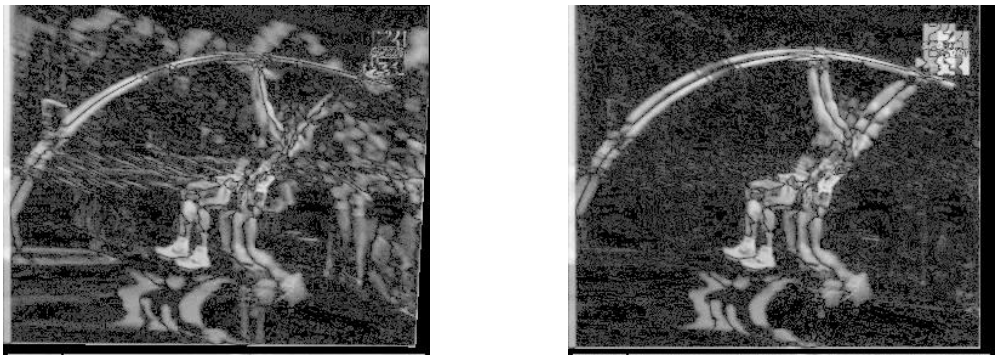
La construction de la MSCV est illustrée Figure 2.24.

a. Deux trames successives.



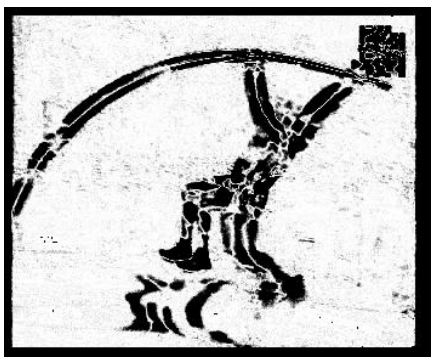
b1. Estimateur non-robuste. b1. Estimateur robuste.

b. Flot optique associé au modèle affine estimé.



c1. Estimateur non-robuste. c2. Estimateur robuste.

c. Visualisation des différences compensées.



d. Visualisation des poids de l'estimateur robuste.

**Figure 2.22.** Estimation robuste et non-robuste du mouvement global pour une scène incluant des objets individuels présentant des mouvements différents de celui de la caméra (vidéo du corpus AGIR).

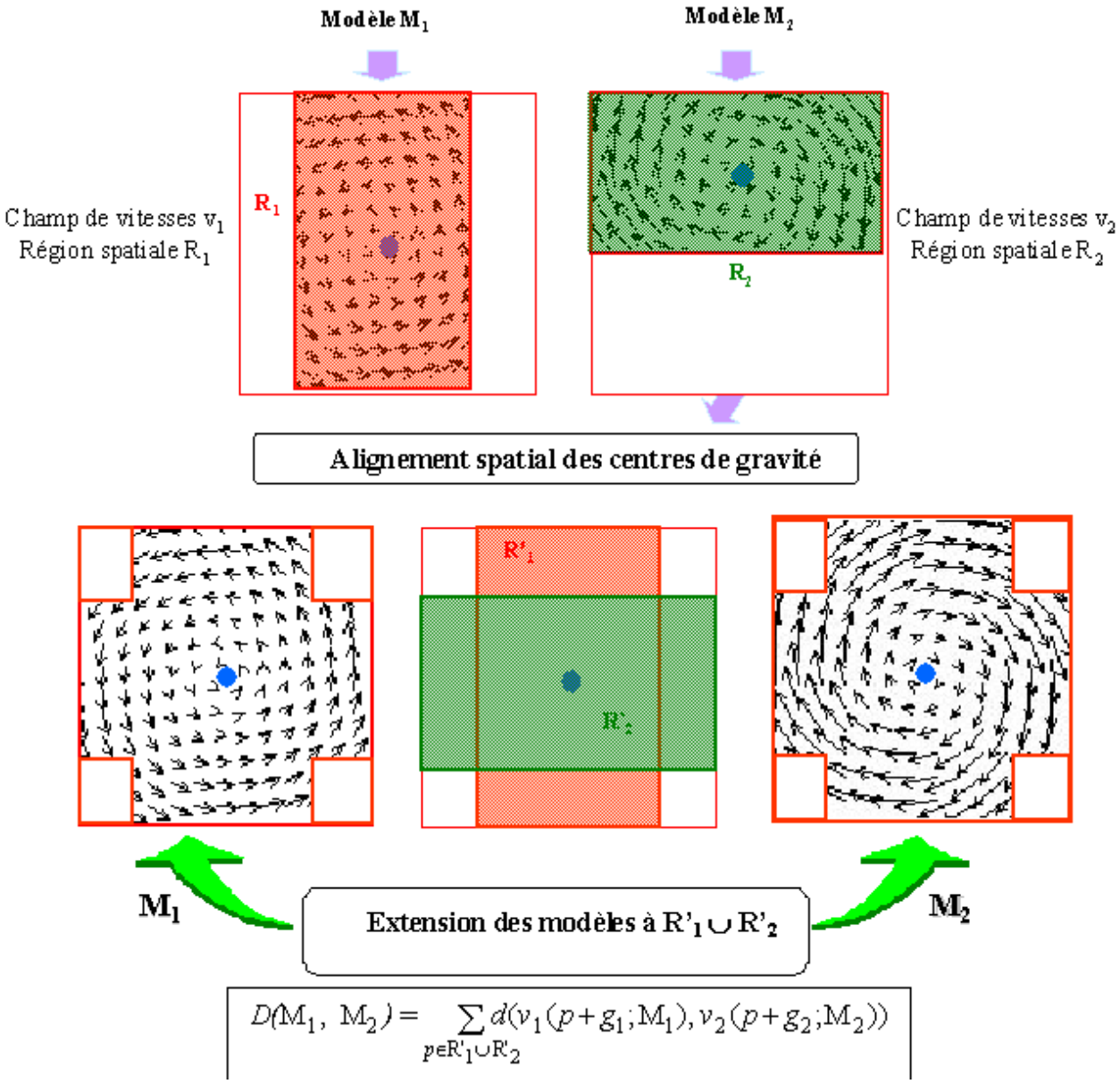
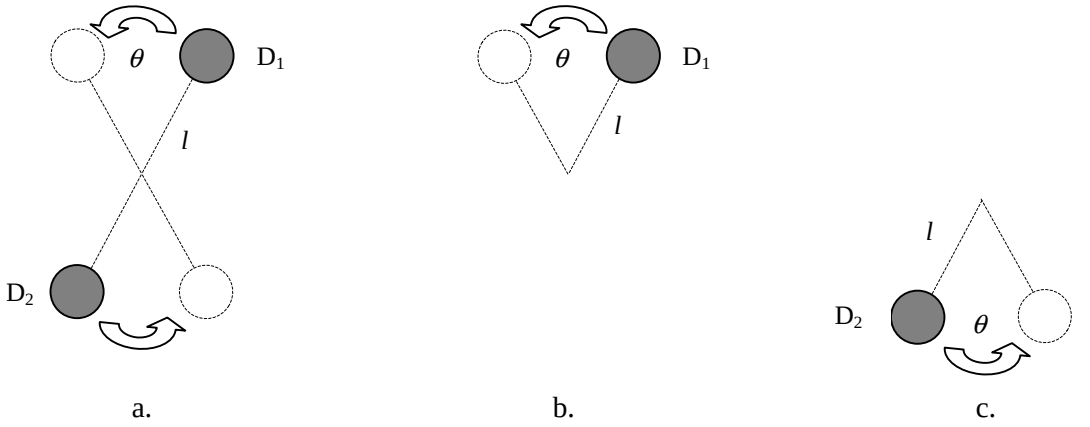



**Figure 2.24.** Construction de la MSCV.

Cette approche permet de prendre en compte la distribution spatiale des vecteurs de vitesse sur les supports des objets considérés. Remarquons que dans le cas des translations pures, le champ de vecteurs vitesse étant constant, les résultats sont équivalents à ceux des MSEP.

Aligner les centres de gravité des objets peut sembler de prime abord un peu arbitraire, puisque cette procédure confère en quelque sorte une signification particulière à la notion de similarité de mouvement. Pour discuter ce point, prenons un exemple simple.

La Figure 2.25 présente deux disques $D_1$ et $D_2$, virtuellement reliés par une tige, cet ensemble étant solidairement animé d'un mouvement de rotation d'angle $\theta$ autour du centre de gravité de cette haltère virtuelle. Son centre de gravité est situé à une distance $l$ du centre de gravité de chaque disque.



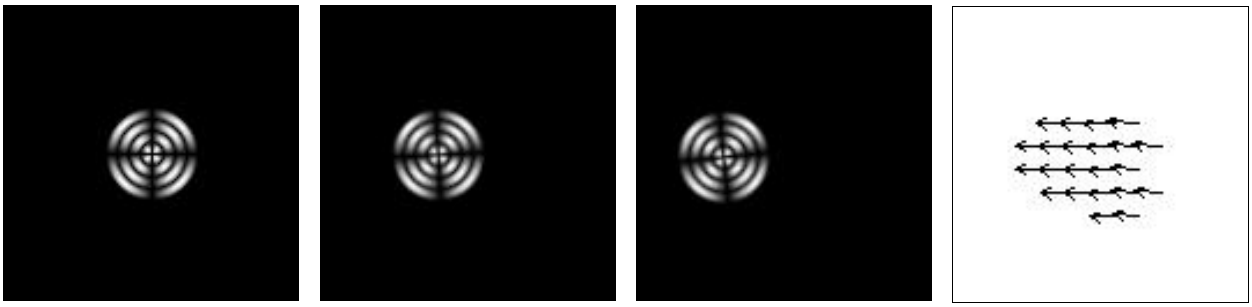

a. b. c.

**Figure 2.25.** Haltère virtuelle dans un mouvement de rotation.

Pour un observateur placé au centre de gravité de l'haltère virtuelle, l'ensemble des deux disques (Figure 2.25.a) et chacune de ses sous-parties (Figure 2.25.b et Figure 2.25.c) présentent alors le même mouvement de rotation.

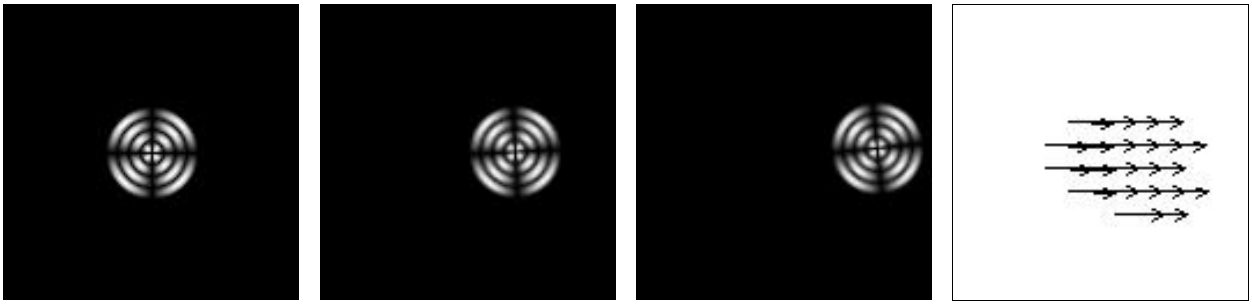
Considérons à présent chacun des deux disques pris séparément et leurs deux séquences synthétiques associés (Figure 2.26), obtenues avec les paramètres suivants :

- angle de rotation $\theta = 3°$,
- distance du centre de gravité de chaque disque au centre de gravité de l'ensemble $l = 5R$, où $R$ désigne le rayon de chaque disque.



$a_1$. Trois trames successives. $a_2$. Champ de vitesse entre les deux premières trames.

a. Disque $D_1$.



$b_1$. Trois trames successives. $b_2$. Champ de vitesse entre les deux premières trames.

b. a. Disque $D_2$.

**Figure 2.26.** Mouvements synthétiques correspondant à chacun de deux disques $D_1$ et $D_2$ et champs de vitesse associés.

Comme nous pouvons le constater, ces objets ont visuellement des mouvements de sens opposés, bien mis en évidence par les champs de vitesse associés. De cet exemple, il ressort que selon le point de vue d'observation adopté les deux disques ont soit le même mouvement, soit des mouvements différents.
Dans ce contexte de paradoxe apparent, quelle stratégie adopter ? Soit une vision globale en considérant le mouvement des objets comme sous-parties d'un système plus large, soit une vision par composante, chaque objet formant alors un "tout" indépendant du reste. Dans la pratique, les objets animé que nous traitons sont considérés comme des entités individuelles et indépendantes. De plus, la caméra est focalisée sur certains objets d'une scène plus large. Cela plaide en faveur d'adopter une d'adopter une stratégie par composante.
Du point de vue de la mesure de similarité sur les mouvements, la stratégie globale revient à aligner les objets par rapport à leurs points critiques (de vitesse nulle), alors que la stratégie par composante consiste à aligner les centres de gravité.  représente un choix, compatible avec la perception visuelle intuitive.
C'est ce choix d'alignement compatible avec la perception visuelle, puisqu'il permet de considérer comme différents des mouvements tels que ceux présentés Figure 2.26, que nous avons donc adopté dans la suite de notre développement. Un raffinement supplémentaire sera défini au Paragraphe 2.5.4, afin de prendre en compte adaptativement les dimensions des objets et les divers types de mouvement.

En fonction des objectifs de la requête, différentes fonctions distance peuvent être considérées (Tableau 2.2).

<table>
<tr><td>*Norme $L_1$ :*<br>$d(v_1, v_2) = \left|v_{1x} - v_{2x}\right| + \left|v_{1y} - v_{2y}\right|,$</td><td>*Fonction de corrélation directionnelle ($C_D$) :*<br>$d(v_1, v_2) = 1 - \frac{1}{2}\left[1 + \frac{\langle v_1, v_2 \rangle}{\|v_1\|_2 \cdot \|v_2\|_2}\right],$</td></tr>
<tr><td>*Norme $L_2$ :*<br>$d(v_1, v_2) = \|v_1 - v_2\|_2 = \left|v_{1x} - v_{2x}\right|^2 + \left|v_{1y} - v_{2y}\right|^2,$</td><td>*Fonction de différence d'amplitudes :*<br>$d(v_1, v_2) = \left|\|v_1\|_2 - \|v_2\|_2\right|,$</td></tr>
<tr><td colspan="2">*Fonction de corrélation non-directionnelle :*<br>$d(v_1, v_2) = 1 - \frac{\left|\langle v_1, v_2 \rangle\right|}{\|v_1\|_2 \cdot \|v_2\|_2},$</td></tr>
<tr><td colspan="2">où $v_1 = (v_{1x}, v_{1y})^\tau$ et $v_2 = (v_{2x}, v_{2y})^\tau$ sont deux vecteurs 2D arbitraires, $\langle \cdot, \cdot \rangle$ désigne le produit scalaire en $\mathbb{R}^2$, $|\cdot|$ la valeur absolue et $\|\cdot\|_2$ la norme $L_2$.</td></tr>
</table>

**Tableau 2.2.** Différentes fonctions distance entre les vecteurs de vitesse 2D.

Ces différentes fonctions distance permettent d'effectuer dans un contexte unifié aussi bien des requêtes bas niveau où certains champs de vitesse sont recherchés, que des requêtes de plus haut niveau comme par exemple la recherche "d'objets approchant la caméra", ou "d'objets en mouvement en translation vers la droite" ou "d'objets en rotation".
La fonction de corrélation directionnelle est à mettre en œuvre dès lors qu'aussi bien la direction que le sens du mouvement sont jugés pertinents. Notons que cette distance occulte l'information d'amplitude du

mouvement. En revanche, pour chercher des mouvements d'un certain type (par exemple, rotations sans spécification du sens), il est préférable de recourir à la fonction de corrélation non-directionnelle. Les distances $L_1$ ou $L_2$ sont recommandées pour retrouver les mouvements les plus similaires à l'exemple. Enfin, la fonction exploitant les différences d'amplitude est utile lorsque des mouvements d'une amplitude donnée sont recherchés.
L'approche proposée permet également de comparer des mouvements décrits par des modèles de types différents (par exemple, comparer un modèle affine avec un modèle projectif), sans perte d'information due à l'approximation de modèles d'ordre supérieur par des modèles plus simples. Des extensions permettant d'étendre cette approche pour comparer des mouvements paramétriques et des mouvements spécifiés par des champs denses de mouvement, ou même des vecteurs de mouvement extraits de flux MPEG-2 ou MPEG-4 sont facilement envisageables dans ce contexte.

Mentionnons en outre qu'une mesure de similarité similaire, fondée sur une fonction distance $L_2$, a été plus récemment considérée de manière indépendante dans [Yu01], pour des objectifs de requêtes par similarité du mouvement global. Les auteurs effectuent ici des requêtes partielles, localisant les minima des fonctions de similarité obtenues en glissant de trame en trame un signal court donné comme exemple à partir d'une autre séquence plus longue.

Le principal inconvénient de la MSCV réside dans sa grande complexité. En effet, elle nécessite de calculer les vecteurs de vitesse et leurs distances sur un ensemble de pixels qui peut être important, et de connaître les régions support associées à chaque objet. L'accès à ces informations à partir d'un disque dur ralentit d'autant les réponses aux requêtes.

Pour contourner ces limitations d'ordre pratique, nous proposons une version simplifiée de la MSCV. L'idée est de remplacer les régions support des objets par de simples boîtes englobantes et de sous-échantillonner les pixels pris en compte dans le calcul des distances. En remplaçant la réunion des régions $\mathbf{R'_1}$ et $\mathbf{R'_2}$ dans l'équation (2.27) par une grille rectangulaire placée à l'origine du repère avec des pas d'échantillonnage en x et y, notés respectivement $\delta_x$ and $\delta_y$, la mesure de similarité simplifiée fondée sur des champs de vitesses (MSSCV), notée $\Delta$, s'écrit :

$$\Delta(\mathrm{M}_1,\ \mathrm{M}_2) = \frac{1}{N_x N_y} \sum_{i=-\frac{N_x}{2}}^{i=\frac{N_x}{2}} \sum_{j=-\frac{N_y}{2}}^{j=\frac{N_y}{2}} d(v_1(p_{ij}+g_1;\mathrm{M}_1), v_2(p_{ij}+g_2;\mathrm{M}_2)), \tag{2.28}$$

où $p_{ij} = \left(\delta_x \cdot i \quad \delta_y \cdot j\right)^\tau$, $N_x$ (respectivement $N_y$) est le nombre d'éléments de la grille de direction horizontale (respectivement verticale) et $p_{ij}$ le noeud d'indices $(i,j)$.
Les pas d'échantillonnage $\delta_x$ et $\delta_y$ s'expriment en fonction des dimensions $L_x$ and $L_y$ de la grille par $\delta_x = L_x / N_x$ et $\delta_y = L_y / N_y$. Il est souhaitable que les dimensions $L_x$ et $L_y$ de la grille soient comparables aux tailles des objets considérés.

### 2.5.3 MSCV et optimisation de la complexité

Pour minimiser la complexité de calcul, le nombre de points d'échantillonnage, déterminé par $N_x$ et $N_y$, doit être le plus petit possible, tout en conservant la sélectivité de la mesure de similarité.
Pour des raisons de simplicité de calcul, considérons une grille circulaire, paramétrée en coordonnées polaires, et dont l'ensemble des points est noté $\Pi$ :

$$\Pi = \left\{ \rho_{mn} = m\delta \cos(\frac{2\pi n}{N}), m\delta \sin(\frac{2\pi n}{N}) \right\}_{m\in\{0,1,\ldots M-1\},\ n\in\{0,1,\ldots N-1\}} . \tag{2.29}$$

L'analogue de la somme dans l'équation (2.28), pour une distance $L_2$, s'exprime par :

$$\Delta(\mathrm{M}_1,\ \mathrm{M}_2) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left\| v_1(\rho_{mn} + g_1; \mathrm{M}_1) - v_2(\rho_{mn} + g_2; \mathrm{M}_2) \right\|^2 . \tag{2.30}$$

Pour les modèles polynomiaux, l'expression sous la somme s'écrit comme une somme de monômes de type $m^{k+l} \cos^k(\frac{2\pi n}{N}) \sin^l(\frac{2\pi n}{N})$ , avec *k* et *l* entiers positifs.
En sommant ces monômes selon la variable *n*, on obtient soit 0, pour $k + l$ impair, soit une valeur proportionnelle à *N*, pour $k + l$ pair, à condition que l'expression suivante soit vérifiée :

$$\sum_{n=0}^{N-1} \cos[(k+l)_{\max} \frac{2\pi n}{N}] = \sum_{n=0}^{N-1} \sin[(k+l)_{\max} \frac{2\pi n}{N}] = 0 . \tag{2.31}$$

Cette expression est toujours vraie pour $2(k+l)_{\max} \le N$ , *N* étant un nombre pair.
Dans ces conditions, il en résulte que la somme :

$$S_m = \frac{1}{N} \sum_{n=0}^{N-1} \left\| v_1(\rho_{mn} + g_1; \mathrm{M}_1) - v_2(\rho_{mn} + g_2; \mathrm{M}_2) \right\|^2 \tag{2.32}$$

est indépendante de *N* pour tout *m*, et donc que la MSSCV

$$\Delta(\mathrm{M}_1, \mathrm{M}_2) = \frac{1}{M} \sum_{m=0}^{M-1} S_m \tag{2.33}$$

est également indépendante de la valeur de *N*. Dans le cas du modèle affine, la plus petite valeur admissible de *N* est $N_{min} = 4$, tandis que pour le modèle quadratique $N_{min} = 8$.
La valeur de *M* contrôle uniquement la pondération des différents coefficients, la MSSCV $\Delta(\mathrm{M}_1,\ \mathrm{M}_2)$ s'exprimant comme une fonction de différents termes $\delta^k F_k(M)$, avec *k* pair, et $F_k(M) = \sum_{n=1}^{M} n^k$ . Remarquons que $F_k(M) = O(M^{k+1})$ et que, asymptotiquement, pour de grandes valeurs de *M*, $\Delta(\mathrm{M}_1,\ \mathrm{M}_2)$ s'exprime comme une fonction du diamètre de la grille circulaire ne dépendant plus de *M*.
Dans le cas de la modélisation affine, la MSSCV s'écrit comme une distance pondérée des paramètres $(a_i)_{i=1}^6$ :

$$\Delta_{aff}(\mathrm{M}_1,\ \mathrm{M}_2)=(a_1^1-a_1^2)^2+(a_2^1-a_2^2)^2+\delta^2\frac{F_2(M)}{2M}\sum_{i=3}^{6}(a_i^1-a_i^2)^2\,. \tag{2.34}$$

Cela n'est plus vrai dans le cas des modèles d'ordre supérieur, comme le modèle quadratique, où des facteurs de couplage entre les paramètres interviennent. Ainsi, pour un modèle quadratique obtient-on l'expression suivante :

$$\begin{aligned}\Delta_{quad}(\mathrm{M}_1,\ \mathrm{M}_2)=\Delta_{aff}(\mathrm{M}_1^{(a)},\ \mathrm{M}_2^{(a)})+\delta^2\frac{F_2(M)}{M}\left[da_1(da_9+da_{10})+da_2(da_{11}+da_{12})\right]+\\ +\delta^4\frac{F_4(M)}{8M}\left[da_7^2+da_8^2+3(da_9^2+da_{10}^2+da_{11}^2+da_{12}^2)+2(da_9da_{10}+da_{11}da_{12})\right].\end{aligned} \tag{2.35}$$

Cette discussion peut être également appliquée au modèle perspectif en considérant une approximation polynomiale par développement en série de Taylor du terme $1/(1 + a_7\, x + a_8\, y)$. Ainsi, pour une approximation du premier ordre, obtient-on un cas particulier du modèle quadratique et $N_{min} = 8$ est suffisant. Des approximations d'ordre supérieur sont également possibles, nécessitant alors d'augmenter la valeur de $N_{min}$.

Ces développements montrent que l'espace des champs de vitesse offre le bon cadre mathématique pour définir de façon optimale des mesures de similarité au sens du mouvement.

## 2.5.4 Alignement, pondération et profil utilisateur

Introduisons un raffinement supplémentaire au niveau des MSSCV, afin de permettre à l'utilisateur d'introduire une certaine connaissance *a priori* lors de la spécification de ses requêtes. En effet, en pratique, où des mouvements combinés (par exemple une translation vers la droite et un zoom avant) sont à prendre en compte, il peut être utile de pondérer différemment la composante de translation du mouvement par rapport à celle homogène. Dans de telles situations, l'utilisateur doit avoir la possibilité de "biaiser" sa requête en renforçant la composante qu'il considère comme la plus pertinente.
Considérons le cas du modèle affine.
La vitesse *v(p)* d'un point $p=(x\ y)^\tau$ s'exprime comme la somme d'un terme constant (translationnel) $v_t$ et d'un terme homogène $v_{\mathrm{hom}}(p)$ :

$$v(p)=v_t+v_{\mathrm{hom}}(p)\,, \tag{2.36}$$

où $v_t=(a_1\ a_2)^\tau$ et $v_{\mathrm{hom}}(p)=Ap$, avec $A=\begin{pmatrix}a_3 & a_4\\ a_5 & a_6\end{pmatrix}$.

Notre but est de définir une stratégie de pondération des deux composantes $v_t$ et $v_{\mathrm{hom}}(p)$, adaptée à chaque type de mouvement. Les mouvements purs, homogènes ou constants, doivent rester inchangés puisqu'une seule composante est présente. En revanche, dans le cas de mouvements combinés, l'influence de la composante homogène peut être renforcée, sans pour autant altérer drastiquement le champ de vecteurs associé au mouvement total, mais tout en diminuant le terme de translation $v_t$. Notons qu'en réduisant $v_t$, nous limitons aussi les problèmes d'alignement spatio-temporel qui peuvent intervenir

comme illustré Figure 2.27.

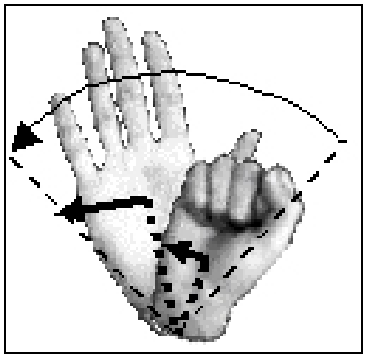


**Figure 2.27.** Le problème d'alignement spatio-temporel. Deux configurations différentes de main subissant la même rotation et dont les successions de trames consécutives sont prises à deux moments différents. Les flèches indiquent les vitesses des centres de gravité, qui sont effectivement différentes.

Il est également possible de considérer la stratégie inverse, qui consiste à pénaliser la composante $v_{\text{hom}}$. Mais comme le plus souvent, dans les mouvements présents dans les vidéos réelles, cette composante est déjà de quelques ordres de grandeur plus petite que la composante de translation, nous ne retiendrons pas cette variante.

Pondérer de manière adaptative, pour chaque mouvement considéré, chacune des deux composantes, revient à définir une mesure de contribution sur celles-ci. Pour cela, nous procédons à une analyse statistique des champs de vitesse sur la grille d'échantillonnage. Notons par $\mu$ et $\sigma$, respectivement la moyenne et la matrice de covariance du champ de mouvement sur la grille considérée :

$$\mu = \frac{1}{N_x N_y} \sum_{i=-\frac{N_x}{2}}^{\frac{Nx}{2}} \sum_{j=-\frac{N_y}{2}}^{\frac{N_y}{2}} v^{ij} \ , \ \sigma = \sqrt{\frac{1}{N_x N_y} \sum_{i=-\frac{N_x}{2}}^{\frac{Nx}{2}} \sum_{j=-\frac{N_y}{2}}^{\frac{N_y}{2}} (v^{ij} - \mu)(v^{ij} - \mu)^{\tau}} \ , \tag{2.37}$$

où $v^{ij} = \left(v_x^{ij} \ v_y^{ij}\right)^{\tau}$ est le vecteur vitesse correspondant au noeud $p_{ij}$ de la grille d'échantillonnage.

Les valeurs propres $\lambda_1$ et $\lambda_2$ de la matrice de covariance sont des valeurs réelles et positives ($\sigma$ étant une matrice positive semi-définie). Elles fournissent une mesure de la variation du champ de vitesse dans les deux directions principales. Notons que dans le cas affine, ces valeurs propres sont proportionnelles à celles de la matrice $A^{\tau}A$. Soit $\lambda_{\max} = \max\{\lambda_1, \lambda_2\}$ la plus grande des deux valeurs propres. Nous définissons donc le rapport $\alpha$ entre mouvement homogène et le terme constant par :

$$\alpha = \frac{\sqrt{\lambda_{\max}}}{\varepsilon + \|\mu\|_2}, \tag{2.38}$$

où $\varepsilon$ est une valeur positive suffisamment petite. Nous utilisons le rapport $\alpha$ pour ajuster de manière adaptative la composante de translation de $v_t$ à $v_t'$, comme décrit dans l'équation (2.39) :

$$v_t' = \frac{v_t}{1 + k\alpha}\,. \tag{2.39}$$

Ici, $k \geq 0$ désigne un facteur de pondération spécifié par l'utilisateur.
Notons que les mouvements purs ou quasi-purs ne seront pas affectés par cet ajustement. En effet, dans le cas de translations, $\alpha$ vaut 0. Les mouvements homogènes ne seront pas plus affectés puisque la valeur de $v_t$ est déjà nulle. Pour des mouvements combinés, de plus fortes valeurs du paramètre $k$ vont diminuer d'autant l'influence de la composante de translation dans le mouvement total.

Une version simplifiée des mesures de similarité fondées sur les champs de vitesses est maintenant disponible pour analyser les performances expérimentales des requêtes par similarité du mouvement en tenant compte à la fois des problèmes d'alignement spatio-temporel et de l'influence des composantes homogènes et translationnelles.

### 2.5.5 Création des bases de test

Evaluer subjectivement et quantitativement les performances d'un descripteur de mouvement dans le cadre de requêtes par similarité nécessite de disposer :

1 – d'une base de test suffisamment grande et représentative des divers types de mouvement,

2 – d'une « vérité terrain » qui consiste en une classification préalable de la base au sens du mouvement en classes distinctes, cohérentes du point de vue du modèle de mouvement identifié et bien représentées, *i.e.* de cardinalité significative (pas de catégorie atomisée).

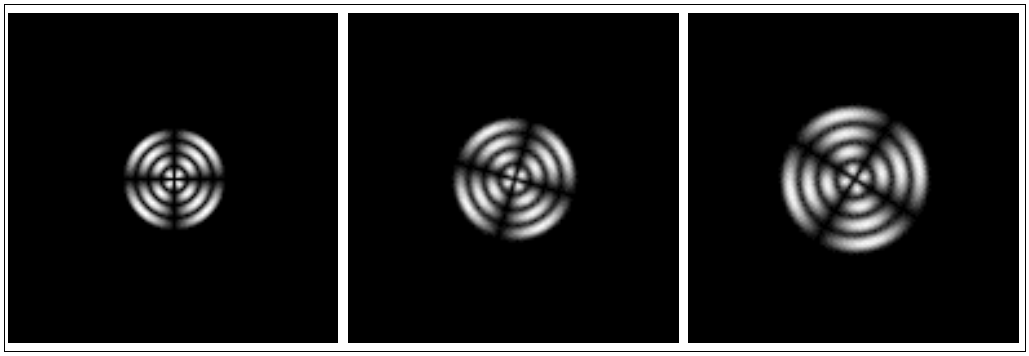
Pour satisfaire ces contraintes, nous avons créé deux bases distinctes, la première composée de 172 mouvements synthétiques et la seconde de 214 séquences vidéos naturelles.
La base synthétique comporte les séquences très courtes (3 trames successives) d'un disque texturé décrivant les mouvements suivants (Figure 2.28) :

- translations pures dans les 8 directions d'une trame carrée en 8 connexité ; pour chaque direction, les composantes horizontales et verticales des vitesses valent successivement 2,5, 5, 7,5 et 10 pixels (soit 32 séquences),
- rotations pures, dans le sens horaire et anti-horaire, de 5, 10, 15 et 20 degrés (soit 8 séquences),
- zooms, avec facteur d'homothétie homogène de 0,8, 0,9 (zoom-arrière), 1,1 et 1,2 (zoom-avant) (soit 4 séquences),
- combinaisons de translations de 5 et 10 pixels dans les 8 directions et rotations à ±5° et ±10° (soit 64 séquences),
- combinaisons de translations, rotations et zooms (translations de 5 pixels dans les 8 directions, rotations des ±5° et ±10°, facteur d'homothétie de 1,1) (soit 32 séquences),
- combinaisons de rotations et zooms (facteur d'homothétie de 0,8, 0,9, 1,1, 1,2 et rotations de ±5°, ±10°, ±15° et ±20°) (soit 32 séquences).

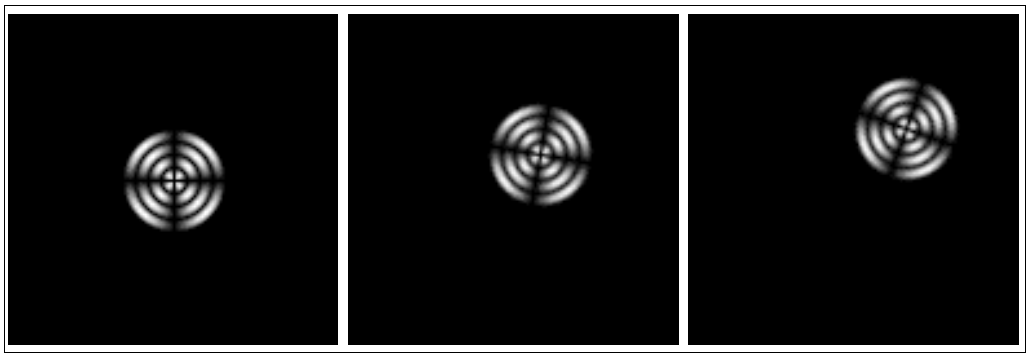
Pour l'ensemble de ces mouvements synthétiques, les paramètres affines de mouvement sont connus *a priori* ce qui dispense de l'étape d'extraction. En outre, les problèmes d'alignement spatio-temporel

n'interviennent pas : la succession de deux trames représentative du mouvement de la séquence est considérée toujours entre la première et la deuxième trame et le support spatial est le même pour toutes les séquences (disque centré au milieu de la première trame). Ces particularités font de la base synthétique une vérité terrain idéale pour l'évaluation du descripteur.

Ici, 12 catégories correspondant aux mouvements purs (translations suivant les 8 directions, rotations dans les deux sens, zooms avant et arrière) ont été définies. Mentionnons que cette base synthétique a été offerte par les auteurs au groupe MPEG-7 pour les évaluations des descripteurs de mouvement du futur standard.



a. Facteur zoom = 1.2, Rotation = 15° (sens horaire).



b. Translation = 10, Direction = NE, Rotation = 10° (sens horaire).

**Figure 2.28.** Deux séquences de la base synthétique.

En pratique, des problèmes d'alignement spatio-temporel des séquences à comparer et de possibles erreurs d'estimation sont à prendre en compte. Afin de valider nos approches dans des conditions plus réalistes, une seconde base de mouvements réels correspondant à différents gestes de la main droite a été réalisée. Le choix de mouvements associés à des gestes naturels est motivé par l'intérêt croissant suscité par les applications de reconnaissance de geste en langue de signe et d'interactions homme-machine dans le cadre d'applications robotiques ou de réalité virtuelle.

Pour différentes configurations de la main droite (paume ouverte doigts écartés, poing fermé, index pointant …), quatre signeurs ont effectué des mouvements de rotation dans le plan de l'image et en 3D, des mouvements de zoom avant et arrière et des mouvements de translation. La Figure 2.29 montre quelques exemples de ces séquences, la main ayant été préalablement segmentée automatiquement selon la méthode décrite au Chapitre 4.

Un ensemble de 214 séquences, chacune d'au plus 40 trames, a été créé et classé en 24 catégories distinctes, décrites dans le Tableau 2.3. Incluant une vérité terrain, cette base permet en outre de tester les

approches développées d'une part en présence de non-alignements spatiaux et d'autre part indépendamment de la forme de l'objet.

Précisons que cette base est une extension de la base de mouvements naturels INT [Zaharia-Iso99e] donnée par les auteurs au groupe MPEG pour évaluer les descripteurs MPEG-7 de mouvement.

Pour chaque séquence, une transition constituée de 2 trames consécutives a été sélectionnée manuellement. Cette transition est supposée être représentative du mouvement de la séquence considérée dans son ensemble. Aucun soin particulier n'a été pris quant au choix de cette succession, située en général vers le milieu des séquences.

Globalement, les deux bases de test constituent un ensemble de 386 séquences à utiliser pour l'évaluation du DMPO.

<table>
<tr><td rowspan="2">Translations Lentes</td><td>N<br>(6)</td><td>NE<br>(8)</td><td>E<br>(21)</td><td>SE<br>(6)</td><td>S<br>(6)</td><td>SO<br>(7)</td><td>O<br>(18)</td><td>NO<br>(9)</td></tr>
<tr><td>Translations Rapides</td><td>N<br>(6)</td><td>NE<br>(4)</td><td>E<br>(6)</td><td>SE<br>(3)</td><td>S<br>(5)</td><td>-</td><td>O<br>(5)</td><td>NO<br>(3)</td></tr>
<tr><td rowspan="2">Rotations</td><td colspan="2">2D</td><td colspan="2">3D verticales</td><td colspan="2">3D horizontales</td></tr>
<tr><td>horaires<br>(16)</td><td>anti-horaires<br>(16)</td><td>gauche-droite<br>(G-D) (7)</td><td>droite-gauche<br>(D-G) (10)</td><td>haut-bas<br>(H-B) (5)</td><td>bas-haut<br>(B-H) (5)</td></tr>
<tr><td>Autres</td><td colspan="2">zooms-avant<br>(9)</td><td colspan="2">zooms-arrière<br>(9)</td><td colspan="2">non-affines<br>(11)</td></tr>
</table>

**Tableau 2.3.** Catégorisation de la base de mouvements naturels
(entre parenthèses - le nombre d'éléments par catégorie).

a. La main est simultanément en translation vers la droite et s'éloigne de la caméra**.**

b. Rotation anti-horaire.

c. Rotation plane horaire combinée à une rotation 3D horizontale du bas en haut.

**Figure 2.29.** Exemples de mouvements extraits de la base naturelle de gestes de la main
(5 trames d'échantillonnage).

### 2.5.6 Résultats expérimentaux

Les expérimentations on été conduites sur les deux bases de mouvement. Les valeurs des différents paramètres intervenant dans le calcul des MSSCV définies dans l'équation (2.28) sont les suivants :

- $\varepsilon = 0.01,\ k \in [0,30]$, $N_x = N_y = 4$, $L_x = L_y = L = 200$ pixels pour les mouvements naturels,
- $\varepsilon = 0.01$, $k = 0$, $N_x = N_y = 4$, $L_x = L_y = L = 80$ pixels pour les séquences synthétiques.

Dans le cadre des expérimentations sur la base synthétique, nous avons tout d'abord considéré les catégories de mouvement pur (*cf.* Paragraphe 2.2) et avons utilisé la mesure de similarité fondée sur la fonction de corrélation directionnelle. Quelle que soit la catégorie prise en compte, tous les autres mouvements de cette même catégorie ont été effectivement retrouvés aux premières places.

Nous avons ensuite analysé les résultats pour des requêtes à partir de mouvements combinés, la MSSCV dépendant alors de la norme $L_2$.

Considérons par exemple une requête combinant rotation anti-horaire et translation vers la droite. La Figure 2.30 présente les cinq premiers résultats classés par ordre décroissant de similarité.

Notons la similitude visuelle des champs de vitesse retrouvés (affichés dans la première cellule de chaque ligne), les mouvements associés correspondant à chaque fois à une translation combinée à une rotation anti-horaire. Ce bon comportement se retrouve pour l'ensemble des mouvements composites.

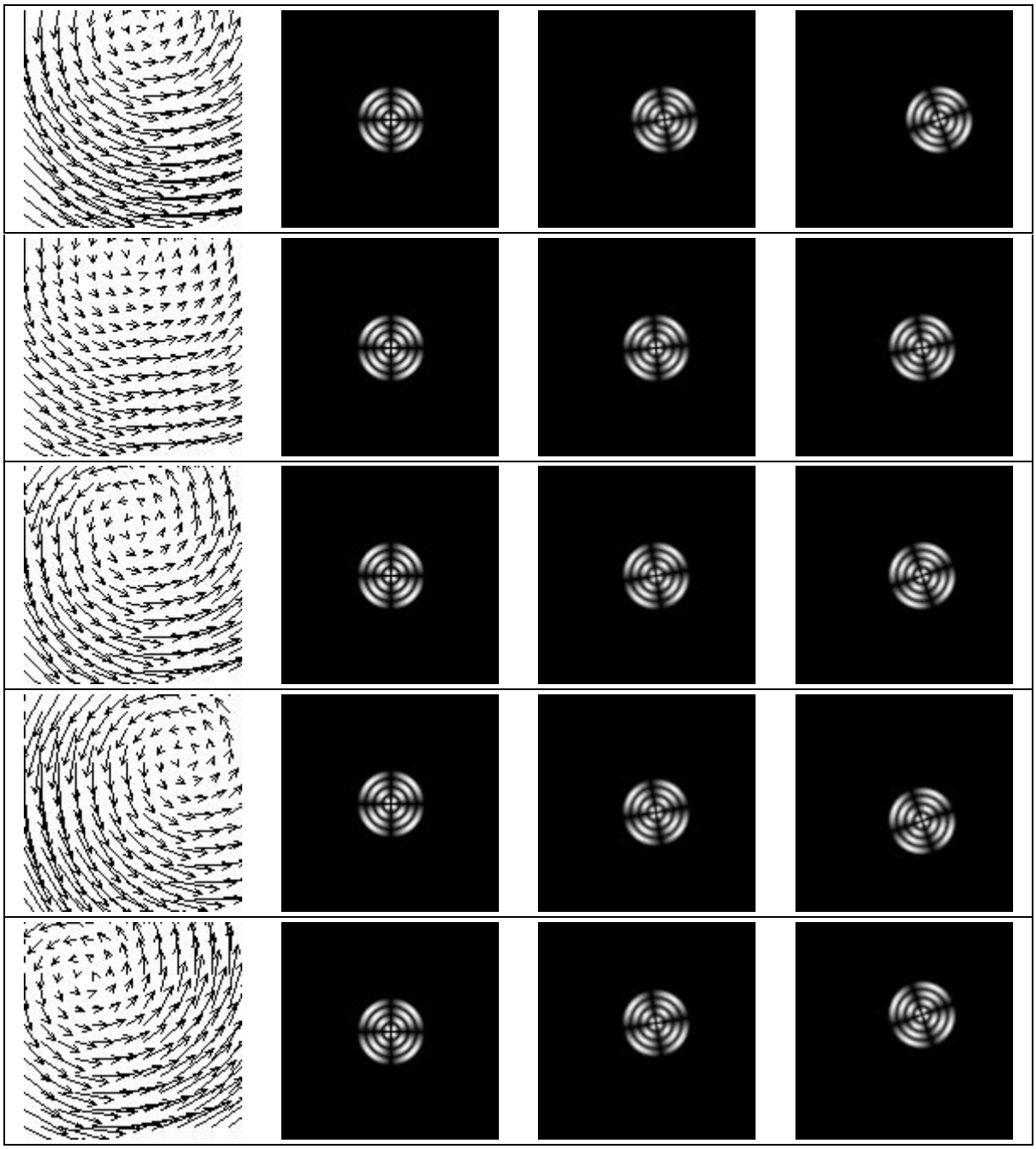


**Figure 2.30.** Résultats d'une requête par similarité sur un mouvement combinant rotation et translation.

Si une catégorie de mouvement est faiblement représentée (comme c'est le cas pour le zoom-avant : 2 séquences seulement), une requête à partir d'un élément de cette catégorie par MSSCV retrouve non seulement aux premières positions les autres mouvement de la catégorie, mais sélectionne ensuite les mouvements composites ayant ce mouvement pur dominant (Figure 2.31). Cela démontre une sorte de "continuité" des MSSCV au sens de la similitude visuelle des mouvements retrouvés par rapport à la requête : l'évolution de ces mouvements lorsqu'on la MSSCV décroît se fait graduellement sans artefacts.

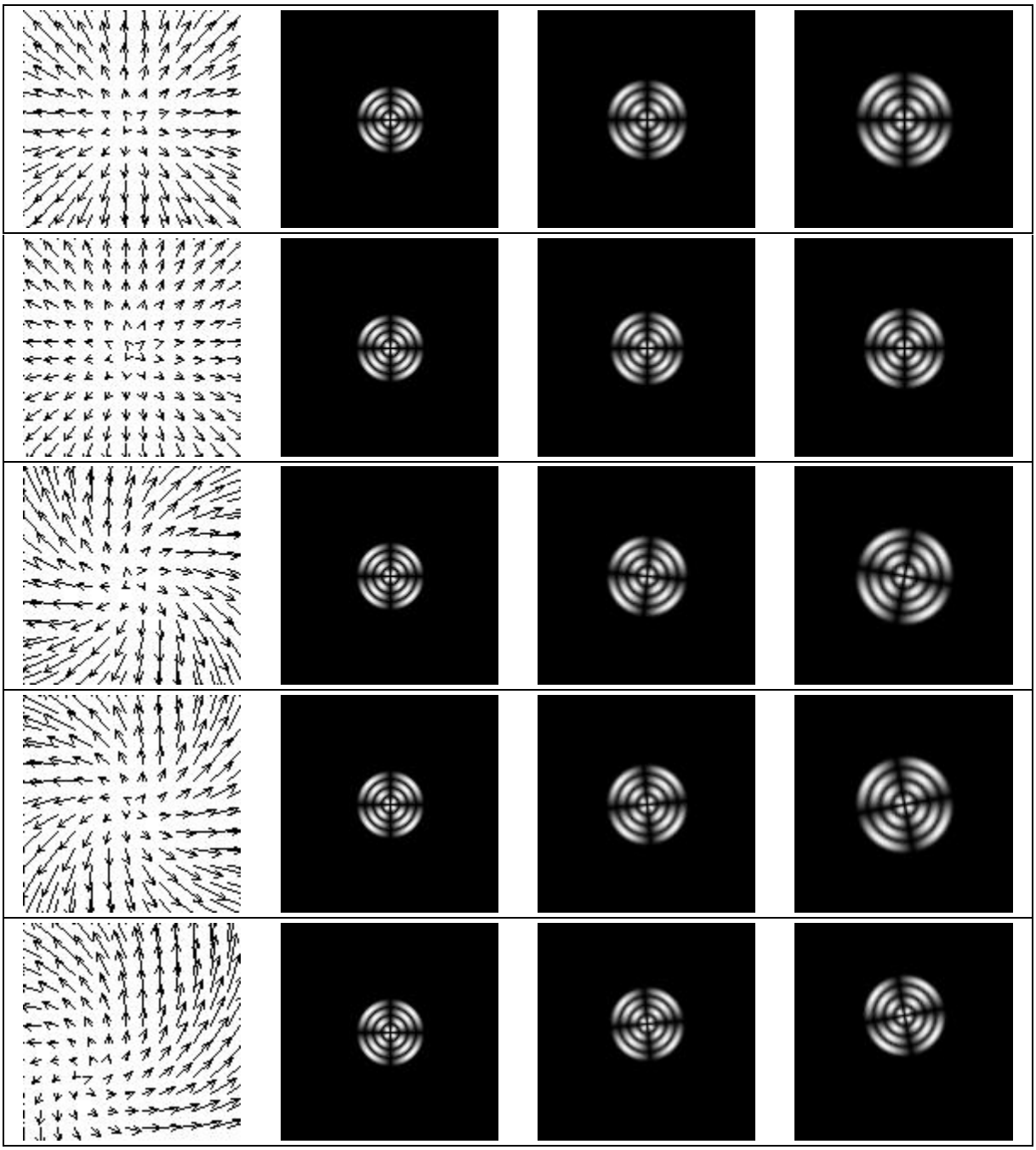


**Figure 2.31.** Résultats d'une requête par similarité correspondant à un zoom-avant.

Ayant ainsi validé notre approche sur la base de mouvements synthétiques, nous avons poursuivi les expérimentations sur la base des mouvements naturels (Figure 2.32 et Figure 2.33).

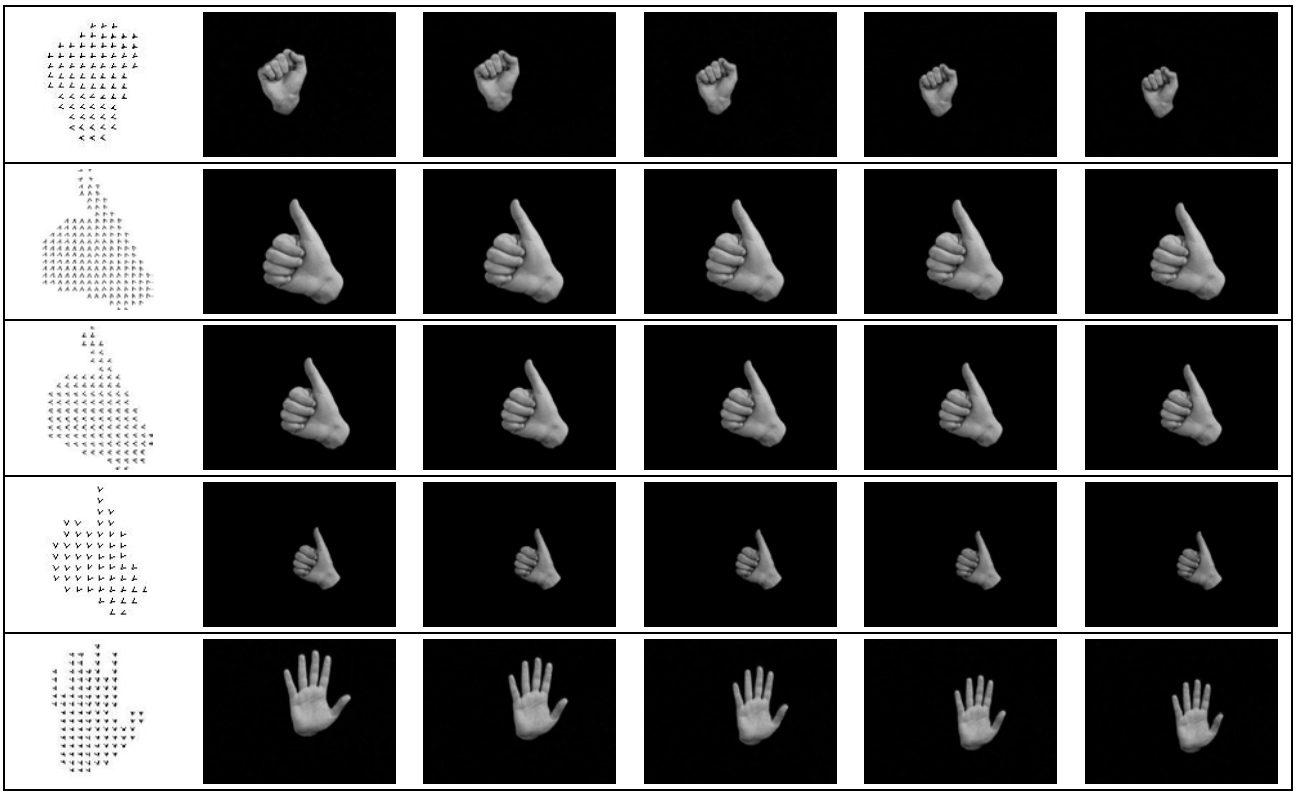


**Figure 2.32.** Résultats de la requête par similarité pour un mouvement zoom-arrière (MSSCV, distance $L_2$, $k$ = 15). Notons la robustesse des mouvements retrouvés par rapport à la configuration et la taille de la main.

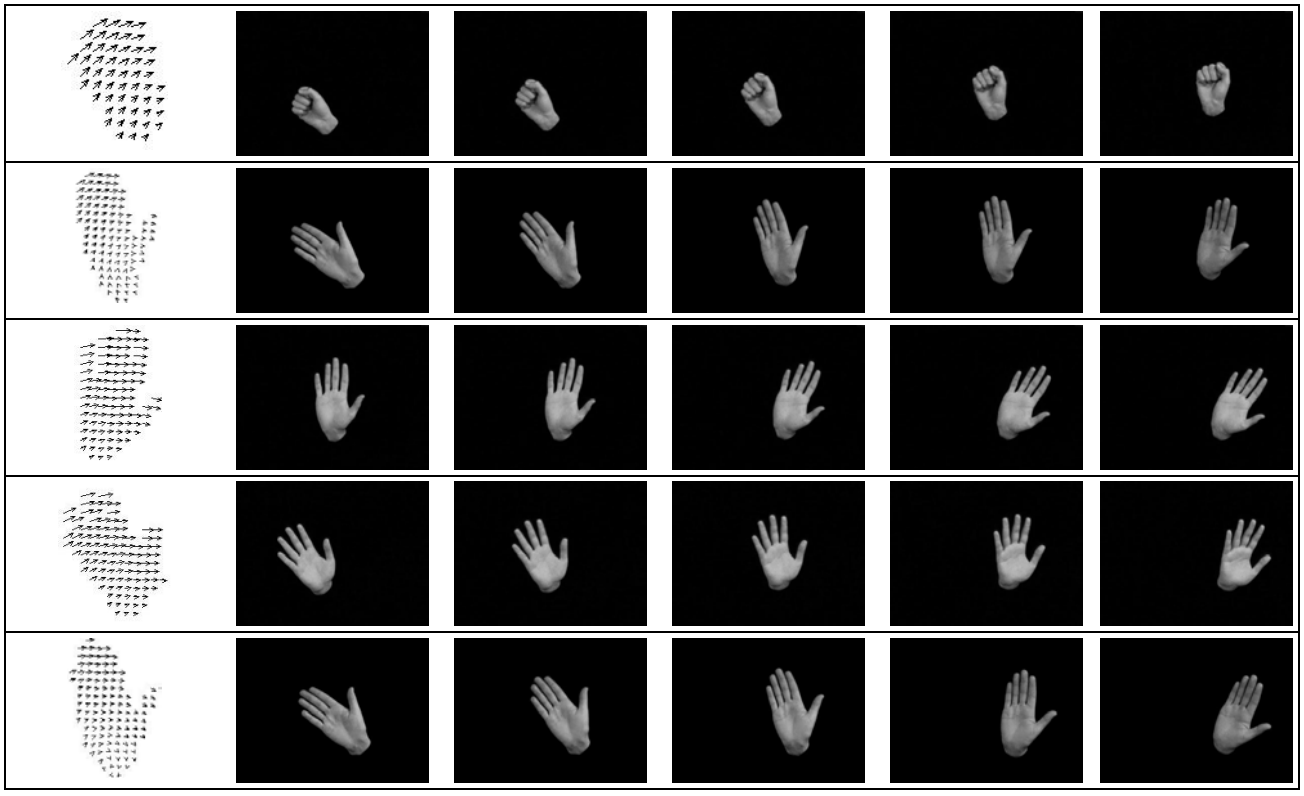


**Figure 2.33.** Résultats de la requête par similarité pour un mouvement de rotation (MSSCV, distance $L_2$, $k$ = 15).

D'une façon générale, les résultats montrent que les mouvements similaires au mouvement de la requête sont retrouvés, malgré les différences de configurations et de dimensions de main et malgré les problèmes d'alignement spatio-temporel bien visibles sur les imagettes de vitesse.
Notons également que comme les composantes de translation sont prédominantes dans la plupart des cas, il est préférable de choisir une valeur de $k$ assez grande ($k = 15$). Mentionnons toutefois la très bonne stabilité des résultats obtenus (quasi-identité) pour $k$ variant de 10 à 30. En revanche, pour $k = 0$, des fausses sélections apparaissent, surtout pour des mouvements dont la composante homogène est très faible par rapport à celle de translation, comme c'est le cas pour les mouvements de zoom. Ainsi, pour la requête de zoom-arrière (Figure 2.32), un zoom-avant est retrouvé en quatrième place lorsque $k = 0$. Cela montre bien la nécessité de faire intervenir le facteur $k$.
Pour des valeurs de $k$ supérieures à 30, les résultats commencent à se dégrader, particulièrement pour les catégories de translation. En effet, ces mouvements sont souvent estimés comme des rotations très faibles dont le centre est très éloigné du centre de gravité de l'objet, par rapport à ses dimensions d'un ordre de grandeur de 5 à 15 fois la taille de la boîte englobante. En augmentant la valeur de $k$, la composante utile de translation est complètement éliminée et les faibles composantes homogènes restant ne sont pas assez pertinentes pour permettre une bonne discrimination.

Pire, les mouvements de translation de sens opposés deviennent confondus, l'information d'orientation du mouvement disparaissant avec cette même composante de translation.
Cela démontre une fois de plus que les problèmes d'alignement doivent être traités conditionnellement aux dimensions des objets et aux types de mouvement pour éviter de confondre des mouvements qui sont qualitativement du même type mais visuellement opposés. D'où l'intérêt et l'utilité du mécanisme adaptatif de pondération des composantes proposé, qui permet d'obtenir des résultats stables pour une large plage de valeurs du paramètre.

Pour évaluer quantitativement et effectivement les performances des différentes mesures de similarité proposées, nous avons calculé le score *Bull-Eye* (SBE) [CE-SM99.07] sur toutes les catégories définies dans le Paragraphe 2.5.5. Pour chaque séquence requête catégorisée, le SBE est défini comme le pourcentage de résultats corrects (*i.e.* mouvements appartenant à la même catégorie que celle de la requête) parmi les 2Q premiers résultats retrouvés, où Q est le nombre d'éléments de la catégorie correspondant à la requête. Pour chaque mouvement de chaque catégorie et pour chaque catégorie le SBE est calculé, puis le SBE moyen par catégorie est estimé (Tableau 2.4) pour les mesures de similarité suivantes : MSEP avec distance $L_2$ et MSSCV avec distances $L_1$, $L_2$ et de corrélation directionnelle ($C_D$).

| ***Catégorie*** | ***Q*** | ***SBE moyen (%)*** | | | |
|---|---|---|---|---|---|
| | | *MSEP* | *MSSCV* | | |
| | | $L_2$ | $L_1$ | $L_2$ | $C_D$ |
| Translation N | 6 | 61 | 59 | 61 | 84 |
| Translation NE | 8 | 58 | 52 | 54 | 68 |
| Translation E | 21 | 72 | 77 | 77 | 92 |
| Translation SE | 6 | 61 | 56 | 54 | 67 |
| Translation S | 6 | 72 | 72 | 73 | 88 |
| Translation SO | 7 | 79 | 75 | 77 | 89 |
| Translation O | 18 | 71 | 71 | 71 | 97 |
| Translation NO | 9 | 55 | 60 | 59 | 78 |
| Rotation horaire | 16 | 46 | 84 | 83 | 100 |
| Rotation anti-horaire | 16 | 52 | 67 | 67 | 89 |
| Zoom-avant | 9 | 40 | 83 | 87 | 59 |
| Zoom-arrière | 9 | 35 | 93 | 95 | 64 |
| 3D vertical G-D | 7 | 30 | 57 | 59 | 75 |
| 3D vertical D-G | 10 | 40 | 81 | 81 | 75 |
| 3D horizontal H-B | 5 | 28 | 40 | 47 | 43 |
| 3D horizontal B-H | 5 | 20 | 47 | 43 | 68 |
| Non-affine | 11 | 31 | 36 | 36 | 44 |
| ***SBE Global (%)*** | | ***50*** | ***65*** | ***66*** | ***75*** |

**Tableau 2.4.** SBE moyens par catégorie.

Il ressort globalement que :

- la MSSCV-$C_D$ offre la meilleure sélectivité (SBE global = 75%), celle-ci persistant même pour les catégories présentant une grande variabilité dans l'amplitude du type de mouvement considéré,
- les MSSCV-$L_1$ et $L_2$ ont des comportements similaires en terme de score moyen global (65%, 66%), et
- les performances liées à la MSEP-$L_2$ chutent avec un SBE global de 50%.

De façon plus détaillée, pour les quatre catégories de mouvements 2D non-translationnels (rotations horaire et anti-horaire, zooms avant et arrière) et pour la distance $L_2$, les résultats font apparaître un gain de 40% par rapport à ceux fondés sur la MSEP. Cela confirme bien que l'espace des champs de vitesse est bien l'espace approprié pour discriminer des mouvements correctement représentés par une modélisation paramétrique 2D.

Pour les deux catégories de zooms, les performances de la MSSCV-$C_D$ (59 et 64%) sont dégradés par rapport à celles obtenues par MSSCV-$L_1$ et $L_2$ (respectivement 83 et 93% et 87 et 95%). Ce résultat, apparemment paradoxal, s'explique par la conjonction d'une part des caractéristiques très spécifiques de ces mouvements, et d'autre part de l'incapacité de la distance $C_D$ à prendre en compte les informations d'amplitude de mouvement, contrairement aux distances $L_1$ et $L_2$. En effet, les mouvements de zoom ont la particularité d'avoir une composante homogène très faible et qui reste faible après la mise en œuvre de la procédure de pondération. La conséquence est que seule la composante de translation détermine les orientations des vecteurs vitesse qui sont alors biaisées et sur lesquelles la mesure $C_D$ s'appuie directement, d'où le mauvais score obtenu. Les mesures $L_1$ et $L_2$ prennent en compte l'information d'orientation et d'amplitude, cette dernière devenant prépondérante en terme de pertinence de

discrimination, d'où les bons scores obtenus.
Cette incapacité de la distance $C_D$ à prendre en compte l'information d'amplitude apparaît de façon évidente dès qu'il s'agit de discriminer parmi une famille de mouvements ceux de faible amplitude de ceux de forte amplitude. A titre d'exemple, nous avons subdivisé chaque catégorie translation en deux classes, translations rapides et translations lentes, et calculé le SBE pour chacune d'elles. Pour la MSSCV-$L_2$, le score est de 77% alors que pour la MSSCV-$C_D$, il descend à 71%. Dans certaines applications comme la reconnaissance de geste en langue de signes, il est utile de pouvoir discriminer des gestes de même type mais effectués plus ou moins rapidement. Dans ce cas, il convient d'utiliser la MSSCV-$L_2$.
Pour les catégories de mouvements 3D, les scores obtenus par la MSSCV sont également significativement supérieurs à ceux donnés par la MSEP. Cette amélioration est d'autant plus marquée que l'estimation du mouvement reflète correctement le mouvement 3D considéré, comme par exemple dans le cas des mouvements verticaux. En revanche, dès que le mouvement est mal estimé (mouvements 3D horizontaux et mouvements non-affines), ce qui est dû à la non-adaptation des modèles paramétriques 2D pour le représenter, le score chute significativement. C'est donc une limite de la modélisation qui est mise en évidence dans le cas des mouvements complexes 3D.
Comme attendu, pour les catégories de translation, les résultats des MSEP et MSSCV sont quasi-équivalents.
En résumé, l'analyse des résultats quantitatifs présentés montre que les MSSCV améliorent de 10 à 25% les performances des requêtes par similarité du mouvement par rapport aux MSEP. De plus, la pondération adaptative développée permet de maîtriser les problèmes d'alignement spatio-temporel et d'influence des composantes translationnelles. Enfin, la définition générique des MSSCV offre un cadre unifié et souple capable d'intégrer plusieurs types de requêtes, adaptés aux préférences de l'utilisateur.

## 2.6 Conclusion

Identifiant l'importance cruciale de disposer de mesures de similarités discriminantes et optimales en terme de complexité de calcul dans le cadre de requêtes par similarité de mouvement, nous avons défini une famille de MSCV dépendant d'une fonction distance générique à spécifier en fonction du type de requête. Les problèmes d'alignement spatio-temporel et de pondération des composantes translationnelle et homogène de mouvement sont analysés et une solution mathématique proposée et mise en œuvre sous forme d'une famille MSSCV et optimisée en terme de complexité de calcul.
Les résultats obtenus démontrent objectivement et quantitativement par estimation du critère *Bull Eye*, la nette supériorité des MSSCV par rapport aux MSEP et établit ainsi la pertinence du DMPO proposé.

# Chapitre 3

# Indexation de maillages 3D par descripteur de forme


## Résumé

*Ce chapitre traite de l'indexation d'objets 3D maillés à l'aide de descripteurs de forme (DF), sous contraintes d'invariance géométrique et de robustesse topologique.*
*Le spectre de forme 3D (SF3D), proposé par les auteurs et retenu comme DF dans MPEG-7, est tout d'abord introduit. Le SF3D est défini comme la distribution d'un index de forme caractérisant localement la géométrie d'une surface 3D et est exprimé comme la coordonnée angulaire de la représentation polaire du vecteur de courbures principales. Intrinsèquement invariant aux transformations géométriques, le SF3D n'est pas robuste vis-à-vis des multiples représentations topologiques d'un même objet.*
*C'est pourquoi un nouveau DF, intrinsèquement stable topologiquement, est proposé. Dérivé de la transformée de Hough 3D, le descripteur de Hough 3D (DH3D) n'est en revanche pas invariant aux transformées géométriques. Nous montrons mathématiquement comment il peut être associé de façon optimale en termes de compacité de représentation et de complexité de calcul à une procédure d'alignement spatial lui conférant alors un comportement d'invariance géométrique. Cela conduit à définir le DH3D optimal (DH3DO).*
*Après avoir spécifié les mesures de similarité utilisées lors des applications de requête et avoir décrit la base de 1300 modèles utilisée, les deux DF sont évalués et comparés objectivement en terme de score Bull-Eye, ce critère établissant une nette supériorité du DH3DO.*



### Mots Clef

*MPEG-7, VRML, maillages 3D, descripteur de forme, invariance géométrique et topologique, requêtes par similarité de forme, courbures principales, spectre de forme, schémas de subdivision, transformée de Hough 3D.*

## 3.1 Contexte et état de l'art

L'indexation d'objets 3D maillés, omniprésents en réalité virtuelle, dans le monde des jeux et en conception assistée par ordinateur, reste encore peu explorée du point de vue des requêtes par similarité de forme en raison des nombreuses difficultés géométriques, topologiques et sémantiques à surmonter.

D'une façon générale, les données 3D se présentent sous plusieurs formes depuis les simples nuages de points, jusqu'aux images de profondeur, en passant par les données volumiques voxelisées comme celles acquises en imagerie médicale, les maillages de la réalité virtuelle et du monde du multimédia interactif. Ici, nous nous intéresserons uniquement aux données 3D, représentées sous forme de maillages polygonaux dont nous rappelons les principales caractéristiques au paragraphe suivant.

### 3.1.1 Maillages 3D

Un maillage 3D (Figure 3.1) est défini par un ensemble de sommets et un ensemble de facettes polygonales. Les positions des sommets des polygones dans l'espace 3D sont exprimées par leurs coordonnées dans un repère cartésien. Cette information spécifie la *géométrie* du maillage, *i.e.* les propriétés métriques de la surface maillée. Les facettes sont définies comme des séquences ordonnées d'indices de sommets et précisent la *connexité* du maillage, *i.e.* les propriétés topologiques de la surface. Les facettes peuvent être des polygones quelconques, avec un nombre arbitraire de sommets. Dans le cas spécifique où toutes les facettes d'un maillage sont des triangles, le maillage est dit *triangulaire*. Tout maillage polygonal peut facilement être transformé en un maillage triangulaire par triangulation [Frey99, DeBerg97] de chacune de ses facettes.

Ajoutons que des attributs photométriques, comme la couleur, la texture et le vecteur normal, peuvent également être associés à un maillage. Traitant ici exclusivement du concept de forme, seules les informations de géométrie et de connexité sont considérées.

Parmi les différents standards de représentation et stockage des modèles maillés (LWO, COB, DFX, 3DS, 3DStudioMax), le *Virtual Reality Modelling Language* (VRML) [VRML97] est devenu l'un des formats les plus populaires. Dans VRML, les différents éléments géométriques, topologiques et photométriques sont listés successivement selon un format textuel : liste des coordonnées des sommets, liste des sommets indexés définissant les facettes... Remarquons que ces représentations textuelles conduisent en général à des fichiers de très grande taille pouvant atteindre des dizaines de MegaOctets pour des modèles un peu élaborés, d'où la nécessité d'introduire des méthodes spécifiques de compression de maillages 3D [Taubin98, Curila99, Alliez01].

Cet enjeu en terme de compression des données maillées est d'ailleurs pris en compte dans le contexte de la normalisation internationale ISO et plus précisément dans le cadre du standard MPEG-4 SNHC (*Synthetic and Natural Hybrid Coding*) [MPEG-4], mais ne révèle pas de ce travail, même si nous avons contribué à l'évalution des solutions techniques apportées [Prêteux-Iso98.10, Prêteux-Iso98.12, Prêteux-

Iso99.03, Prêteux-Iso99.07].

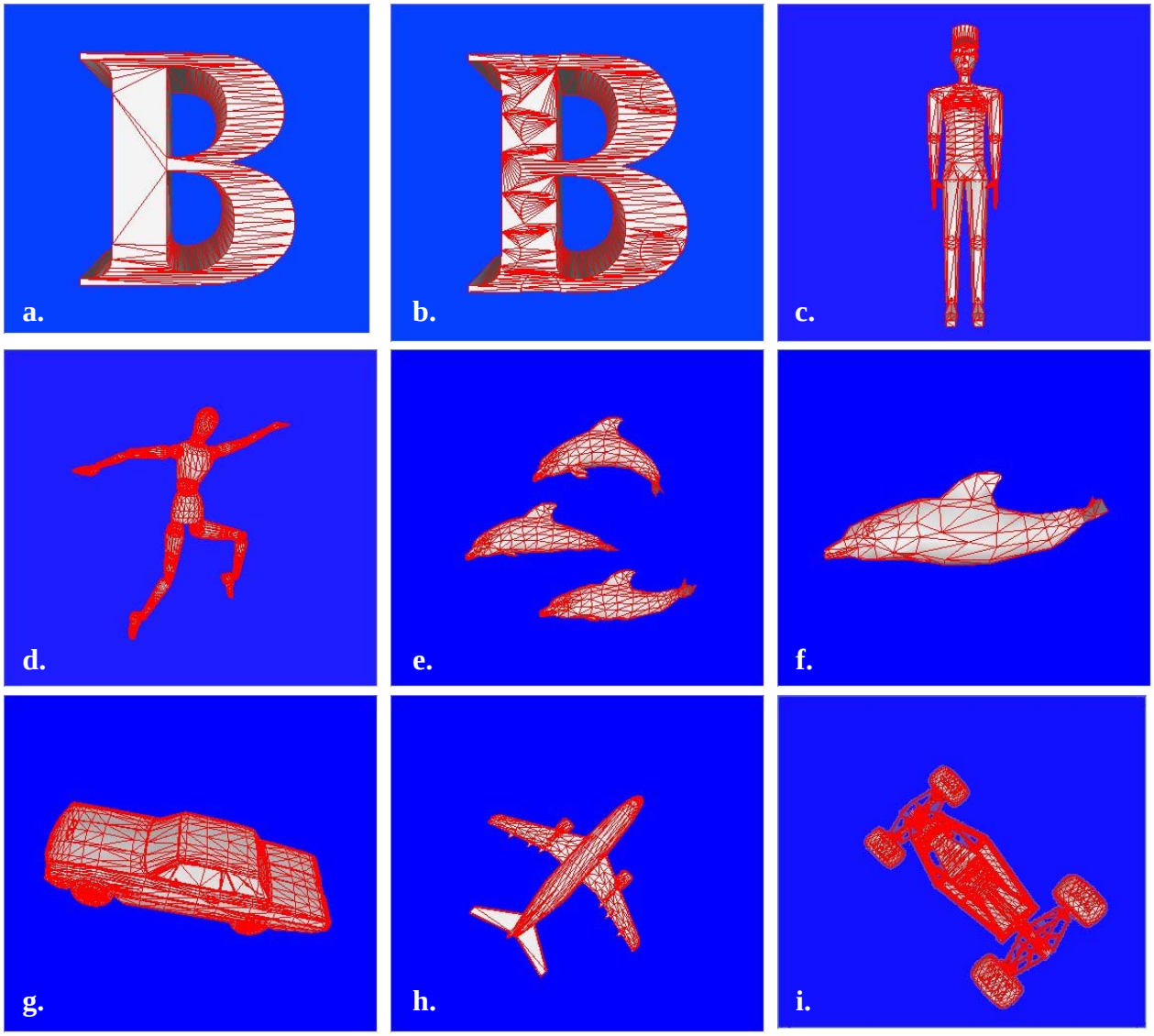



**Figure 3.1.** Exemples de maillages 3D (représentation en fil de fer).

Les représentations maillées relèvent d'une approche purement surfacique des données 3D. Bien que naturel et commode, lorsqu'il s'agit d'application graphiques comme la visualisation et le rendu des surfaces, le caractère intrinsèquement surfacique des données maillées conditionne drastiquement les méthodes d'analyse de forme que l'on peut envisager pour des applications d'indexation et de reconnaissance. C'est l'une des causes des nombreuses difficultés liées à la problématique de représentation automatique par le contenu des maillages 3D. En effet, dans le cas d'objets 2D, les approches "surfaciques" d'analyse de forme se traduisent aisément par des représentations à base de contours 2D, pour lesquels il existe des paramétrisations naturelles, comme par longueur d'arc. De telles paramétrisations s'appliquent aux images 3D de profondeur définies sur des treillis rectangulaires, ce qui en simplifient grandement l'analyse. En revanche, rien de tout cela ne demeure valide dans le cas des maillages 3D.

En outre, notons que même en nous limitant au cadre plus restreint des surfaces fermées, auxquelles il est possible d'associer un volume dual exploitable par les algorithmes d'analyse volumique de forme, un certain nombre de questions persiste en raison de la discrétisation (voxelisation) des données. En effet, pour "voxeliser" de tels maillages, il est nécesssaire de définir la résolution de la grille 3D d'échantillonnage. Cela requiert implicitement de disposer d'une mesure de *l'échelle* des objets, qu'il n'est pas évident de définir de manière cohérente pour tous les modèles d'une base. Les représentations volumiques associées aux maillages sont donc dépendantes de celle-ci. Il ressort de ces observations que les descripteurs de forme pour les maillages 3D doivent prendre en compte la nature surfacique des données maillées.
Ces descripteurs ont à satisfaire à d'autres contraintes liées au facteur humain de l'utilisateur final et au contexte applicatif de l'indexation et de la reconnaissance. Il est en effet nécessaire que ces représentations puissent engendrer des requêtes ou des mécanismes de reconnaissance dont les résultats soient cohérents avec ceux de la perception humaine.
Efin, un descripteur doit à l'évidence vérifier des propriétés géométriques et topologiques plus ou moins intuitives.

Précisons plus avant ces différents aspects.

### 3.1.2 Forme et critères d'invariance

Plusieurs théories de la perception visuelle ont tenté d'appréhender le concept de forme. La psychologie *Gestalt* [Zusne70] se focalise particulièrement sur la perception des projections 2D des objets d'une scène 3D. Dans ses travaux sur la théorie neurophysiologique du comportement, Hebb [Hebb49] soutient la thèse d'une forme distribuée dans des sous-parties de l'objet, d'où le nécessaire apprentissage des relations entre celles-ci pour obtenir des performances de reconnaissance satisfaisantes. Parmi les travaux plus récents et plus proches de la problématique de la vision par l'ordinateur, Koenderink et Van Doorn [Koenderink86] proposent une approche hiérarchique pour décrire l'évolution de la forme à travers différents niveaux de résolution. Des contributions significatives relatives aux différents modules du système visuel humain sont également dues à Marr [Marr76, Marr82] et ont donné naissance aux différentes familles de techniques de reconstruction 3D, appelées génériquement "*Structure from* X" (où X renvoie à du mouvement, de la texture, une silhouette, ou encore de l'ombrage...).

Ces théories visent à définir un contexte générique et établir des critères utiles pour les applications de reconnaissance d'objet en vision par ordinateur. Néanmoins, il n'existe pas de définition unanimement acceptée du concept de forme, couvrant à la fois les multiples aspects de connexité, géométrie, localité et globalité des représentations. En outre, les performances des méthodes de la vision par ordinateur restent encore très pauvres par rapport à celles des systèmes visuels biologiques.

Comme nous le verrons plus loin, la plupart des systèmes de reconnaissance ou d'indexation de forme utilisent principalement des descripteurs de forme (DF) bas-niveau, dépourvus de toute composante sémantique. Or, le concept de forme renvoie implicitement à des informations sémantiques fortes et plurivalentes. Afin d'illustrer concrètement ce point de vue, considérons l'exemple d'un objet articulé,

comme celui d'un *humanoïde*. Ce même concept recouvre en fait une large famille de formes géométriques, dont la variabilité est d'autant plus importante, qu'un même élément peut présenter des apparences bien distinctes, s'agissant ici d'objets articulés et déformables. Les *formes géométriques* correspondant à un homme debout ou assis, sont-elles similaires ? Répondre à cette question, qui renvoie à la théorie de la perception visuelle de Hebb, revient à considérer ou non l'information de position relative des éléments constitutifs de l'objet. Les humains possèdent non seulement la connaissance acquise par l'apprentissage des positions relatives des différentes composantes, mais également la capacité de décider sur la pertinence de cette information, en fonction d'objectifs spécifiques. Ainsi, ils n'éprouvent aucune difficulté pour reconnaître des humanoïdes dans différentes positions ou discriminer les différents gestes effectués par une même personne. En revanche, les ordinateurs sont démunis de pareils mécanismes de décision de haut niveau. Par conséquent, le niveau sémantique à prendre en compte doit être clairement spécifié et les solutions techniques envisagées mises en adéquation avec les objectifs de l'application ciblée. Ainsi, les DF *locaux* (*i.e.*, des DF ne prenant pas en compte les relations spatiales entre les différentes composantes de l'objet, préalablement identifiées) semblent mieux adaptés pour caractériser des humanoïdes en dépit des différentes attitudes sous lesquelles ils sont susceptibles de se présenter, alors que les DF *globaux* (*i.e.*, des DF visant une représentation plus complète, incluant les éléments de localisation spatiale de diverses composantes) apparaissent plus appropriés pour la reconnaissance gestuelle.

Puisque la forme d'un objet 3D est intrinsèquement indépendante de sa position spatiale et de sa taille, un DF doit donc satisfaire à des propriétés d'invariance aux transformations géométriques de similarité (transformations euclidiennes et homothéties). En outre, un objet 3D pouvant être maillé selon de multiples représentations topologiques (Figure 3.1.a et Figure 3.1.b), un DF doit offrir une bonne stabilité par rapport à celles-ci.

Ces considérations précisent le contexte de la représentation d'objets considéré dans ce travail. Comment ces contraintes géométriques et topologiques sont-elles prises en compte par les DF proposés dans la littérature?

### 3.1.3 Synthèse bibliographique

L'indexation par DF des maillages 3D s'inscrit naturellement dans le contexte plus générique de la représentation et la reconnaissance de forme 2D ou 3D.

Les difficultés spécifiques liées à la nature intrinsèquement surfacique des maillages 3D justifient une première catégorie d'approches, dite 2D/3D, dont le principe consiste à associer aux objets 3D un ensemble de projections 2D, correspondant à différents angles de vue. La forme 3D est alors indirectement représentée par des DF 2D quelconques associés à ces images de projection. Notons que ce principe, actuellement intégré dans le standard MPEG-7 (*cf.* Chapitre 1 – *MultiView DS*), permet de réaliser des appariements entre modèles 3D et objets 2D.
Toutefois, pour comparer des objets 3D, ces approches 2D/3D présentent les inconvénients suivants :

- perte d'information due à la projection,
- dépendance des résultats en fonction du nombres de vues,
- absence de méthode robuste pour sélectionner automatiquement le nombre de vues,
- complexité en terme de stockage proportionnelle au nombre de projections,
- temps de réponse aux requêtes en général quadratique avec le nombre de projections.

Pour ces raisons, d'ailleurs consignées objectivement dans le cadre MPEG-7 des évaluations des technologies soumises, nous nous intéressons par la suite à des DF exclusivement 3D.

Les méthodes purement 3D décrites dans la littérature peuvent être répertoriées en approche :

- statistique,
- structurale,
- par transformée,
- variationnelle.

Etudions les méthodes représentatives de chaque catégorie, et soulignons-en avantages et inconvénients.

#### 3.1.3.1 Les approches statistiques

Dans le cadre spécifique de l'indexation, les DF statistiques consistent en général soit à calculer divers moments statistiques [Murao-Iso99.02, Elad00, Zhang01], soit à estimer la distribution de la mesure d'une primitive géométrique donnée (points, cordes, sécantes, triangles, tétraèdres), quelle soit déterministe [Paquet-Iso99, Paquet00, Zhang01] ou aléatoire [Osada01, Elad00].

Dans [Murao-Iso99.02], les auteurs considèrent des moments sphériques, calculés sur un ensemble de points générés à partir du maillage par ré-échantillonnage. Plus précisément, ces points sont obtenus comme intersections des droites reliant le centre de gravité de l'objet aux noeux des grilles carrées obtenues en échantillonnant uniformément chaque facette d'un cube centré au centre de gravité de l'objet. En orientant ces droites, les points sont ensuite classés en *points entrant* dans le maillage et *points sortant* du maillage, les moments sphériques étant calculés pour chaque catégorie.

Dans [Elad00], les moments statistiques sont relatifs aux coordonnées cartésiennes et calculés sur un ensemble de points obtenus en échantillonnant aléatoirement le maillage selon une distribution uniforme sur la surface. Les auteurs proposent ensuite d'utiliser des mécanismes de retour d'usage (*relevance feedback*) [Rui98, Benitez98] à partir des résultats des requêtes pour améliorer itérativement les résultats et créer des profils utilisateurs.

Dans [Zhang01], les divers moments de différents ordres sont calculés à partir d'une représentation volumique à base de tétraèdres définis entre le centre de gravité de l'objet et chaque triangle d'un maillage triangulaire. A partir de quelques formules simples pour calculer les moments volumiques à différents ordres, les auteurs déduisent soit la transformée de Fourier volumique associée à la représentation surfacique de l'objet, soit les moments volumiques globaux.

Utiliser directement les moments statistiques pour la reconnaissance nécessitent une normalisation en taille et position de l'objet, afin d'obtenir une certaine invariance *extrinsèque* de la représentation. Les méthodes de normalisation, qui s'appuient en général sur des mesures globales (comme par boîte englobante, analyse en composantes principales, ...) présentent en revanche un comportement assez instable. Pour cette raison, il est préférable d'utiliser des DF satisfaisant *intrinsèquement* ces propriétés d'invariance.

Les invariants algébriques fournissent des DF globaux, s'exprimant en fonction des moments de différents ordres. La principale difficulté est ici de définir des invariants stables par rapport aux données. De tels invariants et leur mise en œuvre dans le contexte spécifique de la reconnaissance d'objets sont décrits dans [Keren94, Subrahmonia96, Taubin91, Taubin92]. Les mesures de similarité liées aux DF définis comme des vecteurs d'invariants algébriques s'appuient en général sur la distance de Mahalanobis [Mahalanobis36, Lefebvre83], bien adaptée pour prendre en compte les contributions spécifiques de chaque élément composant le DF.

Représentatives des DF par approche surfacique, les EGI (*Extended Gaussian Images*) [Horn84] et leurs variantes [Kang94, Matsuo94], définissent sur la sphère unité (sphère de Gauss) une fonction synthétisant l'information d'orientation en tout point de la surface de l'objet 3D.
Plus précisément, il s'agit ici de construire un histogramme défini sur un ensemble discret d'orientations couvrant la sphère unité. Pour chaque facette d'un maillage, son orientation est quantifiée et l'intervalle de l'histogramme correspondant à celle-ci incrémenté par l'aire (ou par la courbure gaussienne dans d'autres versions) de la facette considérée.
Les EGI classiques caractérisent biunivoquement les objets convexes. L'inconvénient majeur de ces représentations vient de leur grande dépendance à l'information d'orientation. Ainsi, deux objets offrant le même aspect global mais avec des facettes individuelles orientées différemment (comme, par exemple une pyramide en escalier et une pyramide lisse), conduiront à des EGI complètement différentes.

La propriété de complétude de représentation des EGI par rapport aux objets convexes n'est plus valable dans le cas de formes arbitraires non-convexes. Les différentes versions d'EGI plus élaborées se proposent de pallier cet inconvénient en considérant l'information de distance des facettes à l'origine d'un repère préalablement spécifié. Ainsi, dans [Kang94], cette information est-elle prise en compte et intégrée dans la phase d'une fonction complexe, appelée CEGI (*Complex* EGI). En pratique, les intervalles habituels des EGI classiques sont incrémentés d'une valeur complexe, de la forme suivante :

$$Ae^{jd}, \tag{3.1}$$

où $A$ représente l'aire de la facette considérée et $d$ sa distance par rapport à l'origine du repère.

Les auteurs prouvent que les CEGI parviennent à distinguer les objets que les EGI classiques confondent et même démontrent que si les EGI associées à deux objets non-convexes sont identiques, les CEGI correspondantes sont différentes.

Toutefois, les mêmes inconvénients que ceux mentionnés pour les EGI, limitent ces versions améliorées. Les différentes EGI, naturellement invariantes aux translations, ne sont en revanche pas invariantes aux rotations. Cela implique que les méthodes d'appariement à base d'EGI doivent intégrer des mécanismes d'alignement et de mise en correspondance, qui sont en général complexes et coûteux en temps de calcul.

Dans [Paquet-Iso99, Paquet00], les auteurs considèrent comme primitives géométriques la longueur et l'angle des cordes reliant le centre de gravité de l'objet à ses facettes. Ils proposent ensuite plusieurs distributions, plus ou moins complexes, comme par exemple des distributions 1D des longueurs de cordes ou des distributions 2D des angles sphériques de ces cordes (azimut et élévation) dans un repère cartésien précisé.

Dans [Osada01], les primitives considérées sont des segments, triangles ou pyramides, résultant d'un échantillonnage aléatoire des sommets du maillage, les mesures associées étant classiquement celles de longueur, d'aire et de volume. Intrinsèquement invariantes par rapport aux transformations euclidiennes, ces mesures sont normalisées (par exemple par rapport aux valeurs maximale, médiane et moyenne), afin de rendre les différents descripteurs invariants à l'échelle. A retenir ici l'étape préliminaire d'échantillonnage aléatoire, qui permet d'obtenir une représentation de l'objet sous forme d'une famille d'ensembles aléatoires [RandomSets1].

Dans la même catégorie d'approche, mais cette fois dans le contexte spécifique de la reconnaissance de données volumiques 3D issues de l'imagerie en biologie moléculaire, mentionnons l'approche décrite dans [Ankerst99]. Les auteurs proposent différentes partitions de l'espace 3D (concentrique, radiale, ...) pour quantifier les taux de remplissage de l'espace occupé par différentes protéines. Cette approche n'est en revanche pas directement applicable au cas des maillages 3D, en raison de leur caractère surfacique.

Toutes ces descriptions offrent l'avantage de la compacité et d'un faible coût de calcul. Toutefois, trop élémentaires pour caractériser la notion de forme, ces DF ne vérifient pas les contraintes d'invariance géométrique, à l'exception de la méthode proposée dans [Osada01]. Dans tous les autres cas, les auteurs proposent d'introduire un mécanisme d'alignement spatial pour tenter d'y remédier de manière plus ou moins *ad-hoc*, en définissant un repère canonique issu d'une analyse en composantes principales, mais présentant certaines limitations conduisant à des alignements totalement erronés (*cf.* Paragraphe 3.3.2). Quant aux aspects d'invariance topologique, ils sont en général contournés par l'utilisation de représentations auxiliaires ne faisant pas intervenir les caractéristiques topologiques.

#### 3.1.3.2 Les approches structurales

Contrairement aux approches statistiques, les représentations structurales visent à décrire la notion de forme de manière plus complète et plus intuitive. Un premier type d'approche s'appuie sur une segmentation initiale de l'objet en sous-parties satisfaisant certains critères d'homogénéité par rapport à un attribut de forme préétabli, et représentée par des structures spécifiques comme des arbres ou des graphes. Si l'étape de segmentation vise à identifier les différentes structures élémentaires qui composent l'objet considéré, la deuxième phase permet de représenter les relations d'adjacence et éventuellement les positions relatives de ces différentes structures.

Dans une description surfacique d'objets obtenus à partir d'images de profondeur, Dorai et Jain [Dorai97] proposent un DF par graphe de *patchs* maximaux au sens d'un index de forme fonction des courbures principales. L'index de forme (*cf.* Paragraphe 3.2.2) est un critère géométrique différentiel qui permet de classer différentes formes géométriques locales en familles élémentaires comme cylindres, sphères, selles... La connexité de ces *patchs* maximaux est ensuite représentée par un graphe d'adjacence. Enfin, une technique d'appariement de graphes [Ullmann76] permet de gérer la mise en correspondance des surfaces et d'en déduire un critère de similarité morphologique des objets 3D.
Cette approche présente l'avantage de la complétude de la représentation. Les données de profondeur utilisées, bien définies topologiquement sur un treillis rectangulaire, rendent possible l'analyse des relations de connexité des sous-composantes. Toutefois, les méthodes d'appariement de graphes présentent en général l'inconvénient d'une grande complexité de calcul, surtout dans le cas d'objets sophistiqués, composés d'un grand nombre d'éléments.
Une autre technique, également fondée sur l'appariement de graphes est décrite dans [Cordella00]. Il s'agit de détecter différentes sous-composantes de systèmes mécaniques complexes dans le contexte particulier de la conception assistée par ordinateur.

Dans le cadre spécifique de la reconnaissance générique d'objets 3D à partir d'images 2D, Medioni et François [Medioni00] proposent une modélisation par sous-parties volumiques élémentaires des objets 3D à base de cylindres généralisés [Agin76, Ulupinar93, Zerroug94]. Ces éléments, appelés *géons* [Biederman87], sont hiérarchiquement organisés au sein d'un arbre de description, intégrant également des relations d'adjacence. Cette structure hiérarchique permet en outre de définir et de calculer plus efficacement une mesure de similarité fondée sur un coût de transition issu d'une méthode d'appariement par graphes.

Des approches représentatives de la segmentation d'objet 3D en *géons* et des représentations par graphes d'adjacence sont présentées dans [Dickinson92, Dickinson97, Dickinson98]. Les 10 *géons* utilisés ici sont des structures volumiques simples, correspondant à des formes géométriques élémentaires, comme des cylindres (déformés ou non), des cubes, des parallélépipèdes, des cônes et des ellipsoïdes (tronqués ou non), représentés par des maillages 3D simples. Dans [Dickinson91], les géons sont extraits à partir d'images 2D, par mise en correspondance des sous-graphes de primitives d'image avec des sous-graphes des modèles 3D. Cette technique est spécifiquement utilisée dans [Dickinson98] pour l'indexation par le contenu de bases d'images fixes ou de vidéos. L'inconvénient de cette approche réside dans le volume important des interactions requises avec l'utilisateur aussi bien pour la spécification des requêtes que pour l'annotation de la base.
Une approche similaire, mais beaucoup moins élaborée, est décrite dans [Wu94], où les *géons* sont des éléments comme des cubes, ellipsoïdes, cylindres, ayant subi différentes déformations et représentés par des surfaces superquadriques [Terzopoulos91, Pentland91].

Remarquons que les différentes méthodes mettent en œuvre des techniques sous-jacentes d'estimation de surfaces paramétriques (quadriques, superquadriques, cylindres généralisés). Ces techniques d'ajustement seront analysées en détails au Paragraphe 3.2.3.

En général, les approches structurales par graphes de composantes conduisent à des représentations de haut niveau, complètes et élaborées, et en cohérence avec les conclusions des différentes théories de la perception. En outre, elles offrent l'avantage de conduire à des représentations *semi-globales*, exploitables avec différents types d'appariement, aussi bien globaux que locaux. En revanche, elles présentent le handicap de nécessiter la mise en œuvre d'algorithmes d'appariement complexes en temps de calcul, ce qui les rend difficilement exploitables dans le cadre des objectifs spécifiques de l'indexation. Enfin, remarquons que l'analyse de connexité, indispensable lorsqu'il s'agit de déterminer les relations d'adjacence entre les différentes sous-composantes, requiert de disposer de données bien définies topologiquement, comme les images de profondeur considérées par tous les algorithmes mentionnés ci-dessus.

#### 3.1.3.3 Les approches par transformée

Caractéristiques du domaine de la reconnaissance, les représentations d'objets 3D fondées sur des transformées visent à déterminer des représentations de forme globales, définies en terme de transformation intégrale.

Hebert [Hebert95] propose *l'image d'attribut sphérique* (IAS), représentation combinant EGI et un attribut local de forme, l'angle simplexe [Delingette94]. L'IAS est en effet définie comme une fonction de l'angle simplexe sur la sphère unité. En outre, elle caractérise de manière unique les maillages 2-simplexes (*i.e.* maillages pour lesquels chaque sommet a exactement 3 sommets voisins), satisfaisant certaines contraintes de régularité et topologiquement équivalents à une sphère. L'appariement d'objets 3D par des IAS nécessite de trouver la meilleure rotation 3D, au sens d'une fonctionnelle d'énergie définie entre les IAS et conduisant à mettre en œuvre une mesure de similarité.
En pratique, les maillages ne satisfont jamais les contraintes de régularité géométrique et topologique requises par l'IAS. Pour pouvoir appliquer l'IAS à de tels maillages, il serait alors nécessaire de les remailler préalablement en utilisant par exemple des techniques d'ajustement de maillages déformables [Delingette92, Cohen92]. Toutefois, ces techniques sont en général complexes en temps de calcul, impliquant des algorithmes d'optimisation laborieux, et ne sont pas applicables à des objets 3D de topologies arbitraires. En outre, la discrétisation doit être suffisamment dense, afin de garantir la précision de la représentation, ce qui conduit à des descripteurs de taille relativement élevée.

Une approche étroitement liée à la précédente est présentée dans [Zhang99]. Les auteurs utilisent *les images harmoniques* pour représenter des objets 3D topologiquement équivalents à la sphère et convertissent le problème de l'appariement d'objets 3D en un problème d'appariement d'images 2D. L'approche est fondée sur le concept *d'application harmonique* [Xin96], permettant notamment de mettre en correspondance, de manière unique, un domaine 2D avec une surface 3D de topologie sphérique. L'application harmonique est obtenue en pratique par un algorithme de minimisation d'une fonctionnelle énergétique de déformation. Une fois l'application harmonique déterminée, les mêmes attributs locaux de forme que ceux mentionnés dans [Hebert95] sont associés à celle-ci. La reconnaissance est achevée en utilisant une mesure de corrélation entre les images harmoniques.

Une autre approche fondée sur les représentations de surfaces 3D par des images 2D est détailée dans [Johnson99]. Des *imagettes de spin* sont générées pour chaque sommet d'un maillages 3D et utilisées ensuite pour la mise en correspondance selon des critères de corrélation entre les différentes surfaces. Le principal inconvénient réside ici dans la grande complexité de la représentation, qui persiste malgré le recours à une analyse en composantes principales pour la réduire.

Une littérature très riche souligne tout l'intérêt porté aux approches à base de transformée de Hough. La transformée de Hough [Hough62, Illingworth88] s'appuie sur un principe d'accumulation de différentes primitives élémentaires paramétrées selon des variables spécifiques. Les maxima locaux de l'accumulateur indiquent alors la présence de ces différents éléments au niveau des données et permettent leur détection. Initialement appliquée à des droites du plan, paramétrées en coordonnées polaires, la transformée de Hough a été ensuite étendue à d'autres types d'éléments, tels que droites et plans de $\mathbb{R}^3$, coniques, sphères et ellipsoïdes, [Duda72, Tsuji78, Ballard81, Pao92, Hu95, Becker98, Davies98, Bennett99, Guil99, Bhattacharya00].

D'autres axes de recherche considèrent des extensions probabilistes de la transformée de Hough [Xu90, McLaughlin98, Kiryati00], qui visent à réduire la complexité de calcul en échantillonnant aléatoirement les données, sans pour autant diminuer les performances de la représentation en terme de pouvoir de discrimination. Enfin, mentionnons également les versions floues de la transformée de Hough [Bhandarkar94, Philip94, Soodamani98], se proposant de combiner les principes de la transformée de Hough classique avec ceux de la théorie des ensemble flous.
Ces différentes méthodes relèvent d'une approche purement surfacique, consistant à détecter différentes variétés de dimension ($n$-1) plongées dans l'espace $\mathbb{R}^n$. Cependant, remarquons qu'un autre point de vue, cette fois de nature volumique, peut être associé à la transformée de Hough. Cela établit en effet le lien avec la transformée de Radon [Deans83, Gindikin94] qui définit les fondements théoriques de la reconstruction d'objet à partir de ses projections. Comme nous le verrons au Paragraphe 3.3, cette logique nous permettra d'élargir le cadre applicatif de la transformée de Hough au-delà des applications élémentaires de détection de primitives et notamment dans le cadre de l'indexation et des requêtes par similarité.

Toujours dans les approches globales, citons les méthodes spécifiques de la morphologie mathématique [Serra82, Serra88], et notamment les représentations volumiques globales exploitant des notions comme les squelettes [Zhou99, Chuang00, Borgefors99b, Deseiilligny98], les axes médians [Borgefors97, Sheehy96] et les cartes de distance [Borgefors96, Borgefors99a]. Le principe ici est de générer des versions simplifiées, "squelettiques" de l'objet, intégrant toutefois des caractéristiques topologiques et géométriques essentielles. Il est connu que les représentations à base de squelettes sont en général très instables et nécessitent l'introduction d'étapes complémentaires de raffinements et de lissage. Toutefois, ces approches fournissent de puissants outils de représentation, trouvant diverses applications surtout en imagerie médicale. Malheureusement, elles ne sont pas directement applicables aux données surfaciques arbitraires 3D.

De cet ensemble méthodologique plus riche et plus élaboré, il ressort que les DF globaux fournissent des

représentations de forme bien adaptées et quelquefois complètes (au sens où la forme initiale peut être reconstruite à partir de sa représentation). Toutefois, leur absence de robustesse à des déformations locales les rend mal adaptés au caractère générique de la forme dans un contexte d'indexation. En outre, pour garantir leur invariance aux transformations géométriques, il est nécessaire d'introduire des méthodes de normalisation extrinsèque plus ou moins *ad hoc*.

#### 3.1.3.4 Les approches variationnelles

Les approches variationnelles s'appuient sur une modélisation physique des déformations que peut subir une surface 3D donnée. Elles recherchent la solution de l'équation d'équilibre dynamique entre tensions internes et champ de forces externes. La mesure de similarité entre les différentes surfaces est définie comme l'énergie globale de déformation nécessaire afin de mettre en correspondance les surfaces à comparer. Comme représentatifs de cette catégorie d'approche, mentionnons les travaux de [Pentland91, Sclaroff97], où l'analyse des différents modes de déformation permet de définir une mesure de similarité, appelée *appariement modal*.

Ces approches sont particulièrement bien adaptées lorsqu'il s'agit de comparer des modèles appartenant à une même catégorie de formes, comme dans le cas des applications de reconnaissance de visage, pour lesquelles les différences entre les divers éléments sont efficacement modélisées par des déformations non-rigides relativement petites. Une telle approche est appliquée dans [Nastar96] où les images de visage sont représentées comme des surfaces 3D (avec le niveau de luminance comme profondeur) et où l'analyse en composantes principales est utilisée pour l'apprentissage d'un ensemble de déformations représentatives de ladite classe.

En revanche, ces méthodes ne sont plus pertinentes dès lors qu'il s'agit de comparer des objets aussi différents qu'un chat et une bicyclette comme le dit avec humour Jitendra Malik lors de ses conférences [Malik01].

En résumé, l'analyse de la littérature montre l'absence de DF de maillages 3D satisfaisant à la fois les contraintes d'invariance géométrique et topologique. C'est de ce verrou technologique que nous traitons dans ce chapitre.

Identifiant l'enjeu que revêt l'accès aux contenus 3D riches, complexes et protéïformes, le groupe MPEG a ouvert un appel à propositions avec procédures d'évaluation croisée pour lever ce verrou technologique et promouvoir la spécification d'un descripteur de forme 3D dans le cadre du futur standard MPEG-7 [MPEG-7_XM]. C'est dans ce contexte ouvert et normalisé de développements internationaux, présenté au Chapitre 1, que s'inscrivent nos recherches.

### 3.1.4 Les approches proposées

Un premier descripteur, le spectre de forme 3D (SF3D), fondé sur une représentation géométrique locale de surface, que les auteurs ont proposé [Zaharia-Iso99.10s], évalué [Zaharia-Iso99.12s, Zaharia-

Iso00.03b, Zaharia-Iso00.07a, Zaharia-Iso00.07b] et promu dans le cadre du futur standard MPEG-7, est tout d'abord décrit. Intrinsèquement invariant aux transformations géométriques, le SF3D n'est pas robuste vis-à-vis des représentations topologiques multiples.
C'est pourquoi un nouveau DF, le descripteur de Hough 3D (DH3D), intrinsèquement stable topologiquement, est proposé. Dérivé de la transformée de Hough 3D, le DH3D n'est en revanche pas invariant aux transformées géométriques. Nous montrons mathématiquement comment il peut être associé de façon optimale en termes de compacité de représentation et de complexité de calcul à une procédure d'alignement spatial lui conférant alors un comportement d'invariance géométrique. Cela conduit à définir le DH3D optimal (DH3DO).
Après avoir spécifié les mesures de similarité utilisées lors des applications de requête et avoir décrit la base de 1300 modèles utilisée, les deux DF sont évalués et comparés objectivement en terme de score *Bull-Eye*, ce critère établissant une nette supériorité du DH3DO.

## 3.2 Le spectre de forme 3D

Le SF3D [Koenderink90] fournit une représentation intrinsèque d'une surface 3D en exploitant ses caractéristiques géométriques locales. Il est fondé sur la notion d'index de forme qui caractérise localement la forme d'une surface.

Afin de préciser le cadre mathématique nécessaire à nos développements, rappelons tout d'abord quelques éléments de géométrie différentielle des surfaces dans $\mathbb{R}^3$.

### 3.2.1 Eléments de géométrie différentielle

Il s'agit par ces quelques rappels de justifier des propriétés qu'une surface doit satisfaire afin d'en calculer de manière licite des caractéristiques locales comme les courbures de différents types.

#### 3.2.1.1 Surface régulière et paramétrisation

Dans une vision ensembliste, introduisons tout d'abord la notion de surface régulière [DoCarmo76, Spivak79].

**Définition 3.1** (*Surface régulière, paramétrisation*)

Un sous-ensemble $S$ de $\mathbb{R}^3$ est une *surface régulière* si et seulement si pour chaque point $p$ appartenant à $S$, il existe :

- un voisinage $V_p$ de $\mathbb{R}^3$,
- un ensemble ouvert $U_p$ de $\mathbb{R}^2$,
- une fonction $r^p : U_p \subset \mathbb{R}^2 \rightarrow V_p \cap S$,

satisfaisant les conditions suivantes :

(R1) : $r^p$ est une fonction $C^\infty$- différentiable (*i.e.* une fonction admettant des dérivées partielles continues à n'importe quel ordre),

(R2) : $r^p$ est un homéomporphisme (*i.e.* une application bijective et continue, dont la fonction réciproque est également continue),
(R3) : la différentielle de $r^p$, notée $Dr^p$, est inversible.

Si les conditions R1, R2 et R3 sont satisfaites, la fonction $r^p$ est appelée une *paramétrisation* en coordonnées locales du point *p*.

Les conditions R1 à R3 constituent le noyau minimal pour garantir de "bonnes" propriétés mathématiques aux surfaces considérées, comme différentiabilité, invariance par rapport à la paramétrisation, inversion locale et existence et unicité de plans tangents à la surface. Discutons chacune des conditions formulées.

La condition (R1) établit un cadre générique pour le calcul différentiel, l'espace des fonctions indéfiniment différentiables étant fermé par rapport à l'opérateur de dérivée. Mentionnons toutefois que la condition de $C^\infty$- différentiabilité peut être relâchée pour certaines applications bien spécifiques. En ce qui nous concerne, le calcul des courbures s'exprimant comme des dérivées du deuxième ordre, il suffit que la fonction $r_p$ soit $C^2$- différentiable.

La condition de bijectivité imposée par la condition (R2) est nécessaire afin d'éviter des surfaces s'auto-intersectant. Quant aux hypothèses de continuité, notons que $r^p$ est implicitement continue, étant différentiable. La condition complémentaire de continuité de sa fonction inverse est nécessaire pour garantir l'indépendance de caractéristiques différentielles de la surface par rapport à la paramétrisation.

Enfin, la condition (R3) est nécessaire pour garantir l'existence d'un plan tangent en tout point de la surface considérée. Un vecteur *v* est dit *tangent* à la surface *S* au point *p*, s'il existe une courbe paramétrée et différentiable, définie d'un intervalle $(-\varepsilon, \varepsilon)$ dans *S*, avec $\varepsilon$ un nombre réel positif, telle que les propriétés suivantes sont satisfaites :

$$c : (-\varepsilon, \varepsilon) \rightarrow S, \qquad c(0) = p \text{ et } c'(0) = v\ , \tag{3.2}$$

où *c'* désigne la dérivée de la fonction *c*.

Autrement dit, *v* est un vecteur tangent à la surface *S* si et seulement s'il existe une courbe passant par *p* et ayant *v* comme vecteur tangent (ou vitesse). On peut en effet démontrer la proposition suivante :

**Proposition 3.1**

Soient $r : U \subset \mathbb{R}^2 \rightarrow S$ une paramétrisation d'une surface régulière *S* et *q* un point de *U*. L'ensemble des vecteurs tangents à *S* au point $r(q)$ coïncide avec l'image de sa différentielle, notée $Dr^q(\mathbb{R}^2)$.

Rappelons à présent la *notation de Monge* pour les dérivées partielles, qui sera utile pour nos développements futurs. Si $\psi$ est une fonction dérivable jusqu'au deuxième ordre, alors on note classiquement :

$$\Psi_x = \frac{\partial \Psi}{\partial x},\ \Psi_y = \frac{\partial \Psi}{\partial y},\ \Psi_{xx} = \frac{\partial^2 \Psi}{\partial x^2},\ \Psi_{xy} = \frac{\partial^2 \Psi}{\partial x \partial y},\ \Psi_{yy} = \frac{\partial^2 \Psi}{\partial y^2}. \tag{3.3}$$

Remarque :

La condition (R3) de bijectivité de la différentielle $Dr^q$ garantit que $Dr^q(\mathbb{R}^2)$ forme bien un sous-espace vectoriel de dimension 2, *i.e.* un plan dans $\mathbb{R}^3$ ayant les vecteurs $r_u$ et $r_v$ comme base. Soit $p = (u, v)$ un point de $U$, alors le plan tangent à la surface $S$ au point $p$, noté $T_p(S)$, s'exprime par :

$$T_p(S) = \left\{ sr_u(u,v) + tr_v(u,v) \quad \middle| \quad \forall (s,t) \in \mathrm{IR}^2 \right\}. \tag{3.4}$$

Les conditions imposées sur la paramétrisation dans la Définition 3.1 portent sur l'ensemble générique des paramétrisations possibles. En pratique, il est parfois utile de considérer des paramétrisations plus spécifiques, comme celle de Monge introduite ci-dessous.

**Définition 3.2** (*Paramétrisation de Monge*)

Une paramétrisation de Monge est une fonction différentiable $r$, définie d'un ouvert $U$ de $\mathbb{R}^2$ et à valeurs dans $\mathbb{R}^3$, de la forme :

$$r : U \subset \mathbb{R}^2 \to \mathbb{R}^3, \qquad \forall (u,v) \in U, \quad r(u,v) = (\, x(u,v), y(u,v), z(u,v)\,)^t = (\, u,v,z(u,v)\,)^t. \tag{3.5}$$

La différentielle $Dr$ de $r$ au point $(u,v)$ de $U$ est une application linéaire définie de $\mathbb{R}^2$ à valeurs dans $\mathbb{R}^3$, et s'exprimant dans les bases canoniques de $\mathbb{R}^2$ et $\mathbb{R}^3$ sous forme de la matrice $(3 \times 2)$ ci-dessous :

$$Dr : \mathbb{R}^2 \to \mathbb{R}^3, \qquad Dr = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \end{pmatrix}^t = \begin{pmatrix} 1 & 0 & \frac{\partial z}{\partial u} \\ 0 & 1 & \frac{\partial z}{\partial v} \end{pmatrix}^t, \tag{3.6}$$

dont le rang est évidemment égal à 2. Il en résulte que la dimension de l'espace image du $\mathbb{R}^2$ par la différentielle $Dr$ est égale à 2 :

$$\dim (\, Dr(\mathbb{R}^2)\,) = 2. \tag{3.7}$$

$Dr$ est donc inversible et par conséquent, l'image de $r$, notée $S = r(U)$, est une surface régulière.

A chaque plan tangent, peut être associé un vecteur normal (orthogonal). Si l'on se limite à l'ensemble des vecteurs normaux unitaires, à chaque plan tangent à $S$ au point $p$, noté $T_p(S)$, on peut associer deux vecteurs normaux, notés $N_p$ et $-N_p$, ayant la même direction mais des sens opposés. Quel que soit le choix de ces vecteurs normaux, ils déterminent de manière unique le plan tangent.
Pour une paramétrisation donnée $r : U \subset \mathbb{R}^2 \to S$ autour d'un point $p$ de la surface considérée $S$, on peut construire un champ de vecteurs normaux différentiable, noté $N$, en prenant :

$$N : r(U) \rightarrow \mathbb{R}^3, \qquad N(q) = \frac{r_u \times r_v}{\|r_u \times r_v\|}(r^{-1}(q)). \tag{3.8}$$

Localement, on peut donc associer au voisinage de chaque point d'une surface régulière, pris conditionnellement à la paramétrisation considérée, un champ différentiable de vecteurs normaux. Toutefois, il n'est pas toujours possible d'associer un champ de vecteurs normaux sur l'ensemble de la surface régulière *S*. Pour cela, il faut que la surface considérée soit *orientable*.

**Définition 3.3** (*Surface régulière orientable*)

Une surface régulière *S* de $\mathbb{R}^3$ est dite orientable si et seulement s'il existe une famille de *patchs* $(V_i)_{i \in I}$ recouvrant *S*, de paramétrisations associées $\varphi_i : U_i \rightarrow V_i$, telle que pour tout $(i, j) \in I^2$ pour lequel $V_i \cap V_j \neq \phi$, le changement de coordonnées $\varphi_{ij} = \varphi_i \circ \varphi_j^{-1}$ est de jacobien positif.

Intuitivement, le concept de surface orientable peut être expliqué comme suit. Dans un voisinage suffisamment petit du point considéré, chaque vecteur normal induit un sens de parcours sur des courbes fermées infinitésimales de la surface. Ce sens peut-être par exemple associé à la normale par la règle du tire-bouchon (Figure 3.2). La surface est orientable si, au voisinage d'un point appartenant à l'intersection de deux *patchs* distincts, les sens de ces courbes dans les deux patchs coïncident.

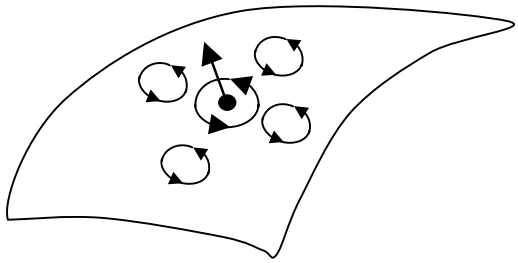
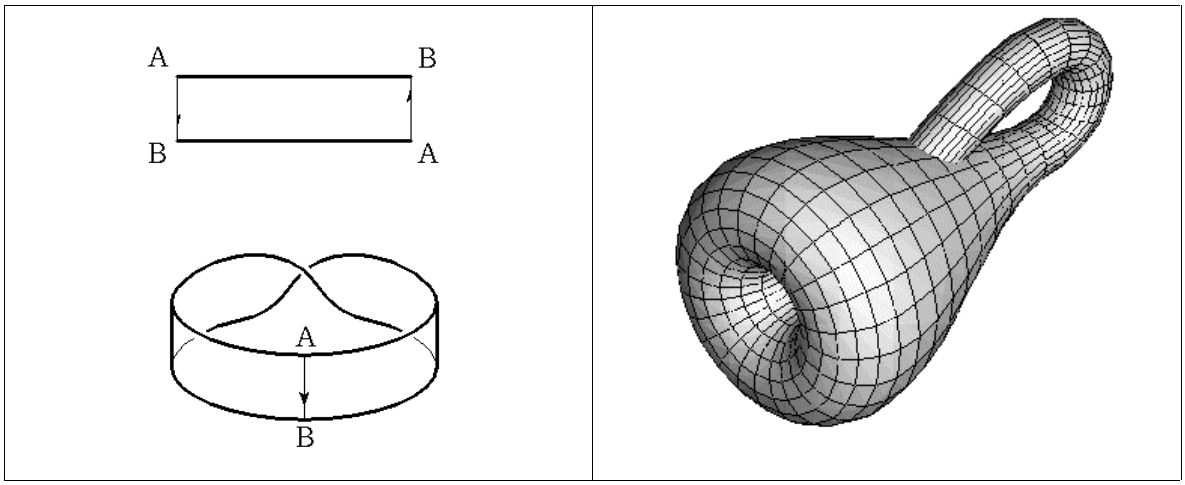


**Figure 3.2.** Exemples de courbes infinitésimales associées aux vecteurs normaux.

Nous verrons par la suite que cette interprétation géométrique sera particulièrement utile pour définir la notion d'orientation d'une surface maillée (*cf.* Paragraphe 3.2.3).

Il existe néanmoins des surfaces pathologiques, qui, bien qu'étant régulières, ne sont pas orientables, comme les célèbres ruban de Möbius et bouteille de Klein (Figure 3.3).




a. Ruban de Moëbius. b. Bouteille de Klein.

Figure 3.3. Surfaces non-orientables.

Le lien entre concept d'orientation d'une surface et propriétés du champ de ses vecteurs normaux est établi dans la proposition suivante.

**Proposition 3.2**

Une surface régulière $S$ de $\mathbb{R}^3$ est dite orientable si et seulement s'il existe un champ de vecteurs normaux $N : S \to \mathbb{R}^3$, continu. Le champ $N$ sera alors également différentiable et appelé *orientation* de la surface $S$.

En considérant uniquement des vecteurs normaux de longueur unité, on obtient un champ de vecteurs particulier, connu sous le nom d'application de Gauss.

**Définition 3.4** (*Application de Gauss*)

Soit $S$ une surface régulière avec une orientation $N$. L'application $N : S \to S^2 \subset \mathbb{R}^3$, où $S^2 = \left\{ (x, y, z) \in \mathrm{IR}^3 \,\middle|\, x^2 + y^2 + z^2 = 1 \right\}$ désigne la sphère unité de $\mathbb{R}^3$, est appelée application de Gauss.

La sphère unité de $\mathbb{R}^3$ sera désormais appelée *sphère de Gauss*.

Intéressons-nous maintenant à quelques propriétés métriques des surfaces, comme celles de longueur et de courbure.

### 3.2.1.2 Formes fondamentales et courbures

La *première forme fondamentale* est définie dans un contexte riemannien comme une forme quadratique définie positive, exprimant localement les longueurs infinitésimales des différentes courbes de la surface (Figure 3.4). La première forme fondamentale en un point $p$ de $S$, notée $I_p$, s'exprime comme la restriction du produit scalaire de $\mathbb{R}^3$ au plan tangent $T_p(S)$ :

$$\begin{aligned} I_p : T_p(S) &\rightarrow \mathbb{R}, \\ w &\mapsto \langle w, w \rangle = \|w\|^2 . \end{aligned} \tag{3.9}$$

En considérant la base $(r_u, r_v)$ de $T_p(S)$ associée à la paramétrisation considérée, quel que soit un vecteur arbitraire $w$ de $T_p(S)$, il existe des coordonnées réelles $\alpha_u$ et $\alpha_v$, telles que $w$ s'exprime dans cette base sous la forme :

$$w = \alpha_u r_u + \alpha_v r_v . \tag{3.10}$$

La première forme fondamentale s'exprime alors en coordonnées locales par :

$$I_p(w) = \langle w, w \rangle = \left\langle \alpha_u r_u + \alpha_v r_v , \alpha_u r_u + \alpha_v r_v \right\rangle = \left\langle r_u , r_u \right\rangle a_u^2 + 2\left\langle r_u , r_v \right\rangle a_u a_v + \left\langle r_v , r_v \right\rangle a_v^2 . \tag{3.11}$$

En notant :

$$F = \left\langle r_u , r_v \right\rangle, \quad E = \left\langle r_u , r_u \right\rangle \quad \text{et} \quad G = \left\langle r_v , r_v \right\rangle, \tag{3.12}$$

nous obtenons alors :

$$I_p(w) = E a_u^2 + 2F a_u a_v + G a_v^2 = \begin{pmatrix} a_u & a_v \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix} \begin{pmatrix} a_u \\ a_v \end{pmatrix}. \tag{3.13}$$

En particulier, considérons une courbe paramétrée de la surface, passant par $p$ :

$$\forall\, \varepsilon > 0,\ \forall\, t \in (-\varepsilon, \varepsilon), \qquad \alpha(t) = r(u(t), v(t)), \tag{3.14}$$

avec $\alpha(0) = p$.

Sa dérivée en $t = 0$, notée $\alpha'(0)$, représente la vitesse de la courbe et est un vecteur de $T_p(S)$. En lui appliquant la première forme différentielle, on obtient :

$$I_p(\alpha'(0)) = \left\langle \alpha'(0), \alpha'(0) \right\rangle = E u'^2(0) + 2F u'(0) v'(0) + G v'^2(0) . \tag{3.15}$$

La première forme différentielle permet donc d'exprimer les longueurs infinitésimales des différentes courbes de la surface dans le repère local.

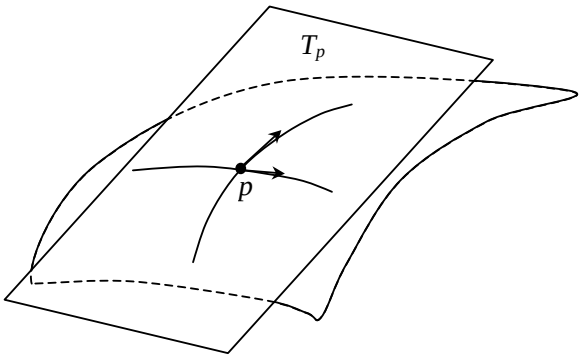



**Figure 3.4.** La première forme fondamentale donne les vitesses des différentes courbes sur $S$. Les normes de ces vitesses représentent les longueurs infinitésimales des courbes.

Quant à la *deuxième forme fondamentale*, elle exprime localement la flexion de la surface selon différentes directions, à partir de la différentielle du champ d'orientation $DN_p : T_p(S) \rightarrow T_p(S)$. Rappelons tout d'abord que $DN_p$ est un endomorphisme symétrique, comme décrit dans l'équation suivante :

$$\forall\, w_1, w_2 \in T_p(S), \quad \left\langle DN_p(w_1), w_2 \right\rangle = \left\langle w_1, DN_p(w_2) \right\rangle . \tag{3.16}$$

On peut donc lui associer une forme quadratique, appelée deuxième forme différentielle et définie comme suit.

**Définition 3.5** (*Deuxième forme différentielle*)

Soient S une surface régulière et orientable et *p* un point de *S*.

La forme quadratique $II_p$ définie sur $T_p(S)$ par le produit scalaire suivant :

$$II_p(v) = \left\langle DN_p(v), v \right\rangle, \tag{3.17}$$

est appelée deuxième forme différentielle de *S* au point *p*.

Interprétons cette formulation (Figure 3.5). Soient *p* un point de *S*, *v* un vecteur de $T_p(S)$ et $N_p$ le vecteur normal à *S* en *p*. L'intersection de *S* avec le plan $(v, N_p)$ s'appelle *une section normale* de *S* au point *p* et définit une courbe plane *c*, dont la normale *n* est donnée par $\pm N_p$, selon la paramétrisation considérée. En outre, pour une paramétrisation de *c* par longueur de courbe, la courbure de *c* au point *p* est égale en valeur absolue à $II_p(v)$.

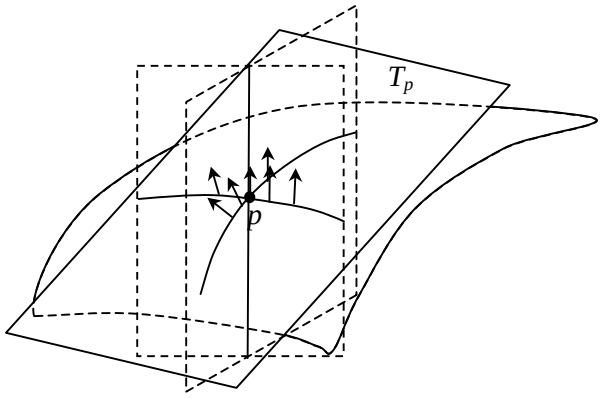



**Figure 3.5.** La deuxième forme fondamentale donne les courbures des sections normales de *S*.

Remarquons qu'étant une forme quadratique, $II_p$ est diagonalisable par un changement approprié de repère.

Ses valeurs propres, désignées par $k_p^1$ et $k_p^2$, sont appelées *courbures principales*. Les vecteurs propres associés donnent les directions principales de variation des courbures des sections normales. En outre, si $k_p^1 \neq k_p^2$, alors les directions principales sont orthogonales. On peut donc construire une base orthonormée $\{e_1, e_2\}$ dans le plan tangent $T_p(S)$, avec $e_1$ et $e_2$ vecteurs propres unitaires.

Mentionnons également le caractère extrémal des courbures principales. Si l'on considère dans le repère orthonormé $\{e_1, e_2\}$ un vecteur *v* de $T_p(S)$, de longueur unitaire et de direction $\theta$, donné par :

$$v(\theta) = e_1 \cos(\theta) + e_2 \sin(\theta), \tag{3.18}$$

la deuxième forme différentielle s'exprime alors comme une fonction de $\theta$ :

$$k_n(\theta) = II_p(v(\theta)) = k_p^1 \cos^2(\theta) + k_p^2 \sin^2(\theta), \tag{3.19}$$

dont les *extrema* sur l'intervalle [*0, 2π*) sont donnés par les courbures principales $k_p^1$ et $k_p^2$. Cela justifie la terminologie de courbure *minimale* et *maximale* associée à $\min\{k_p^1, k_p^2\}$ et $\max\{k_p^1, k_p^2\}$, respectivement.
L'équation (3.19) est connue sous le nom de *théorème d'Euler*.

En partant de l'équation (3.19), le lieu géométrique des points ($\xi$, $\eta$) du plan tangent satisfaisant la relation $k_p^1 \xi^2 + k_p^1 \eta^2 = \pm 1$ est une courbe, appelée *indicatrice de Dupin*. Intuitivement, cette courbe est similaire à celle obtenue en "coupant" la surface avec un plan parallèle à $T_p(S)$ et à une hauteur par rapport à ce dernier de $\pm\varepsilon$, où $\varepsilon$ est un nombre réel positif suffisamment petit. L'indicatrice de Dupin peut être une ellipse, une hyperbole ou deux droites parallèles. Elle induit une classification de la forme locale des surfaces en elliptique, hyperbolique et parabolique, respectivement. Nous verrons par la suite comment l'index de forme, attribut de surface utilisé pour décrire localement la forme d'une surface, est étroitement lié à l'indicatrice de Dupin.

Enfin, remarquons que pour une base de $T_p(S)$ fixée, on peut exprimer $DN_p$ sous forme d'une matrice (2×2) notée :

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \tag{3.20}$$

où les coefficients sont des réels.
Cette matrice n'est symétrique que dans une base orthogonale. Le déterminant et la trace de la matrice *A* sont en revanche invariants par rapport au choix de la base et nous permettent d'introduire les notions suivantes :

**Définition 3.6** (*Courbure moyenne, courbure gaussienne*)

> Soient *S* une surface régulière orientable, *p* un point de *S* et $DN_p : T_p(S) \rightarrow T_p(S)$ la différentielle de l'application de Gauss en *p*. Le déterminant et la demi-trace de $DN_p$ sont respectivement appelés *courbure gaussienne* de *S* en *p* (notée *K*), et *courbure moyenne* de *S* au point *p* (notée *H*).

En choisissant la base orthonormée définie par les vecteurs propres ($e_1$, $e_2$) de la deuxième forme fondamentale, les courbures gaussienne et moyenne s'expriment en fonction des courbures principales comme suit :

$$K = k_1 k_2 \qquad , \qquad H = \frac{k_1 + k_2}{2} \quad . \tag{3.21}$$

Les courbures moyenne et gaussienne jouent un rôle important en analyse différentielle des surfaces géométriques, en raison de leurs remarquables propriétés.
La connaissance des courbures moyenne et gaussienne permet de déduire les valeurs des courbures principales, comme solutions de l'équation du deuxième degré suivante :

$$x^2 - 2Hx + K = 0, \tag{3.22}$$

La courbure gaussienne s'interprète également comme la limite suivante :

$$K = \lim_{aire(V)\to 0} \frac{aire(N_p(V))}{aire(V)}, \tag{3.23}$$

où $V$ est un voisinage de $p$ suffisamment petit.
Cette relation correspond en effet à la définition originale que Gauss a donnée pour la courbure d'une surface. Elle généralise de façon naturelle la notion de courbure d'une courbe au cas d'une surface, la courbure étant alors égale à l'inverse du rayon de la sphère approchant localement et au mieux ladite surface.

En ce qui concerne la courbure moyenne, elle est la moyenne de $k_n(\theta)$ (appelée aussi *courbure normale* dans la direction $\theta$) sur le cercle unité de $T_p(S)$ :

$$H = \frac{1}{2\pi} \int_0^{2\pi} k_n(\theta) d\theta . \tag{3.24}$$

Enfin, mentionnons que la matrice $A$ caractérisant la différentielle $DN_p$ dans le repère local ($r_u$, $r_v$) peut être exprimée en fonction des deux formes fondamentales par :

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} e & f \\ f & g \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1} = II\, I^{-1}, \tag{3.25}$$

où

$$I = \begin{pmatrix} E & F \\ F & G \end{pmatrix} = \begin{pmatrix} \langle r_u, r_u \rangle & \langle r_u, r_v \rangle \\ \langle r_v, r_u \rangle & \langle r_v, r_v \rangle \end{pmatrix} \text{ et } II = \begin{pmatrix} e & f \\ f & g \end{pmatrix} = \begin{pmatrix} \langle r_{uu}, N \rangle & \langle r_{uv}, N \rangle \\ \langle r_{vu}, N \rangle & \langle r_{vv}, N \rangle \end{pmatrix}, \tag{3.26}$$

désignent la première et la deuxième forme fondamentale, exprimées dans le repère local considéré. La relation (3.25) est appelée *équation de Weingarten*. Les valeurs propres de la matrice $A$ sont les deux courbures principales.

Dans le cas d'une paramétrisation de Monge de la surface $S = (x,y,z) = (x,y,f(x,y))$, les deux formes différentielles s'expriment par :

$$I = \begin{pmatrix} \langle S_x, S_x \rangle & \langle S_x, S_y \rangle \\ \langle S_x, S_y \rangle & \langle S_y, S_y \rangle \end{pmatrix} = \begin{pmatrix} 1 + f_x^2 & f_x f_y \\ f_x f_y & 1 + f_y^2 \end{pmatrix}, \tag{3.27}$$

$$II = \begin{pmatrix} \langle S_{xx}, N \rangle & \langle S_{xy}, N \rangle \\ \langle S_{xy}, N \rangle & \langle S_{yy}, N \rangle \end{pmatrix} = \frac{1}{\sqrt{1 + f_x^2 + f_y^2}} \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}, \tag{3.28}$$

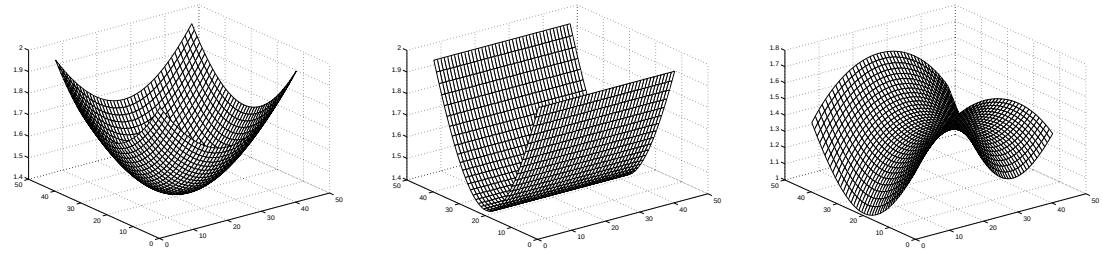
L'équation de Weingarten permet alors d'exprimer très aisément les deux courbures principales en fonction des dérivées partielles du premier et deuxième ordre de la fonction *f*.

## 3.2.2 L'index de forme

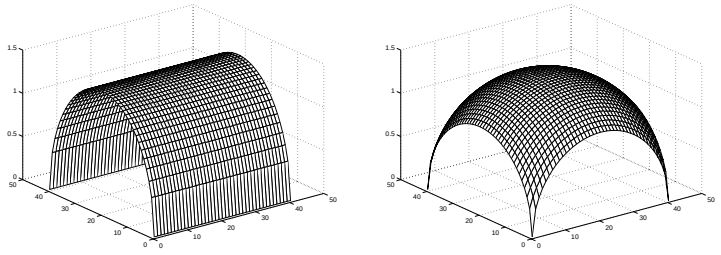
L'index de forme, introduit par Koenderink [Koenderink90], est défini comme la valeur de la coordonnée angulaire de la représentation polaire du vecteur des courbures principales. Soient *p* un point d'une surface 3D $C^2$-différentiable *S* et $k_p^1$ et $k_p^2$, avec $k_p^1 \geq k_p^2$, les courbures principales au point p. L'index de forme s'exprime par :

$$I_p = \frac{1}{2} - \frac{1}{\pi} arctg \frac{k_p^1 + k_p^2}{k_p^1 - k_p^2} \ . \tag{3.29}$$

Il prend donc ses valeurs dans l'intervalle [0,1], n'est pas défini pour les surfaces planes, mais est invariant aux transformations euclidiennes (les courbures l'étant elles aussi) et aux homothéties (étant exprimé comme une fonction du rapport des courbures principales).
L'index de forme permet de caractériser différentes formes élémentaires (Figure 3.6) et a été utilisé pour l'indexation d'images de profondeur [Dorai97] et d'images fixes 2D [Nastar97].



a. Ombilic minimal (IF = 0.0) b. Vallée (IF = 0.25) c Selle (IF = 0.5)



d. Crête (IF = 0.75) e. Ombilic maximal (IF = 1.0)

**Figure 3.6.** Formes élémentaires et leurs indices de forme (IF) associés.

Les rappels de base de géométrie différentielle s'inscrivent dans le contexte mathématique de l'espace continu $\mathbb{R}^3$ et des surfaces 3D régulières. Or, les données 3D considérées en pratique sont des entités intrinsèquement discrètes, puisque définies par des maillages. Si les propriétés de continuité restent alors implicitement vérifiées, celles relatives à la différentiabilité ne sont en revanche jamais satisfaites.

Dans ce contexte de géométrie discrète, est-il possible, malgré tout, de conserver un sens aux éléments différentiels tels que courbures principales et de les estimer de façon fiable et précise ?

## 3.2.3 Le calcul des courbures principales sur un maillage discret

Les différentes approches mises en œuvre se distinguent suivant qu'elles génèrent ou non une surface paramétrique 3D régulière associée au maillage discret considéré.
Ainsi, les méthodes non-paramétriques s'affranchissent-elles de toute référence à une surface 3D régulière en s'inscrivant dans un contexte discret et en effectuant directement le calcul sur le maillage.

### 3.2.3.1 Approches non-paramétriques

Parmi les approches non-paramétriques, citons les approches détaillées dans [Taubin95, Delingette92, Desbrun00].
Desbrun *et al.* [Desbrun00] dérivent les expressions des courbures gaussienne et moyenne en s'appuyant sur quelques théorèmes intégraux. Ainsi, pour calculer la courbure gaussienne, appliquent-ils le théorème de Gauss-Bonnet qui permet notamment d'exprimer la moyenne de la courbure gaussienne sur un domaine triangulaire en fonction de la somme des angles extérieurs sur la frontière du domaine considéré (appelé aussi *angle solide* [Delingette92]). Quant à la courbure moyenne, elle est calculée comme la moyenne du laplacien de la surface sur le même domaine triangulaire. Son calcul s'appuie sur le théorème de Gauss, qui permet d'exprimer l'intégrale surfacique du laplacien comme l'intégrale curviligne de la divergence de la surface sur la frontière du domaine considéré.
Une fois les courbures moyenne et gaussienne déterminées, les valeurs des courbures principales se déduisent facilement à partir de l'équation (3.21).

Delingette [Delingette92] propose une méthode originale pour déterminer la courbure moyenne, dans le cas particulier des maillages 2-simplexes (*i.e.* maillages hexagonaux pour lesquels chaque sommet a exactement 3 sommets voisins), en introduisant la notion d'*angle simplexe*. L'angle simplexe est en effet un équivalent de l'angle solide pour les maillages simplexes et permet d'exprimer la courbure gaussienne. Il est intéressant de noter que l'auteur démontre, en raison de la dualité (maillage simplexe - triangulation) que la courbure moyenne est mieux adaptée pour caractériser les maillages simplexes, tandis que la courbure gaussienne peut être estimée de manière plus fiable pour les triangulations. Il conclut alors que les méthodes d'estimation discrètes ne permettent pas de déterminer à la fois et de manière précise les deux courbures en un même point.

Taubin [Taubin95] utilise également une relation intégrale qui lui permet de déterminer une matrice dont les vecteurs propres sont les mêmes que ceux de l'opérateur $DN_p$ et les valeurs propres associées sont en

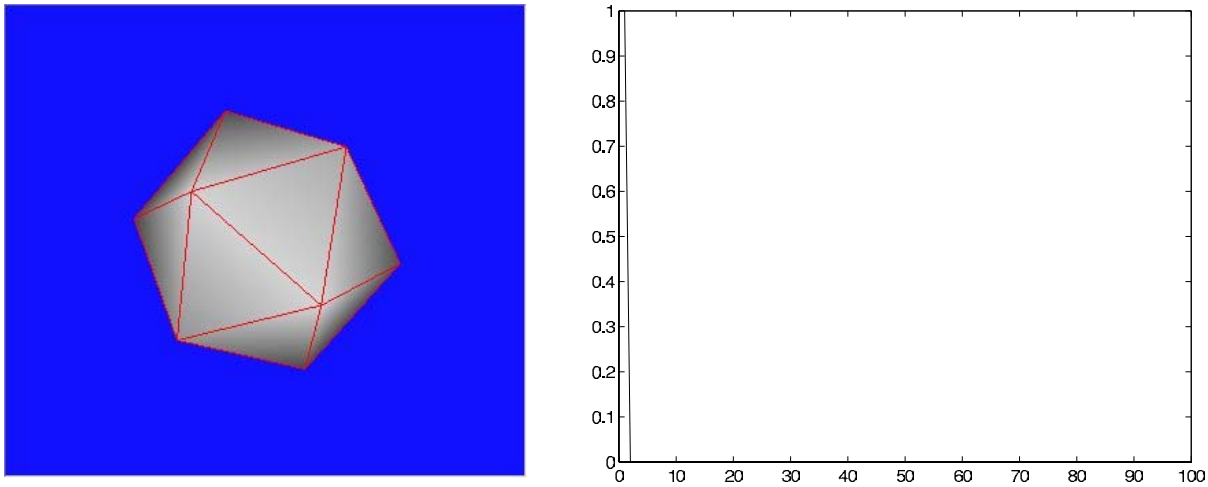
correspondance biunivoque avec les courbures principales. La construction de cette matrice fait intervenir les courbures directionnelles, selon les différentes directions associées aux arêtes reliant le sommet considéré à ses voisins. Ces courbures sont alors estimées par différences finies, en réalisant un développement en série de Taylor au deuxième ordre des sections normales associées à ces différentes directions.

Dans tous les cas, les méthodes non-paramétriques, supposent implicitement que les hypothèses suivantes, portant sur la nature de la discrétisation de la surface, sont satisfaites.

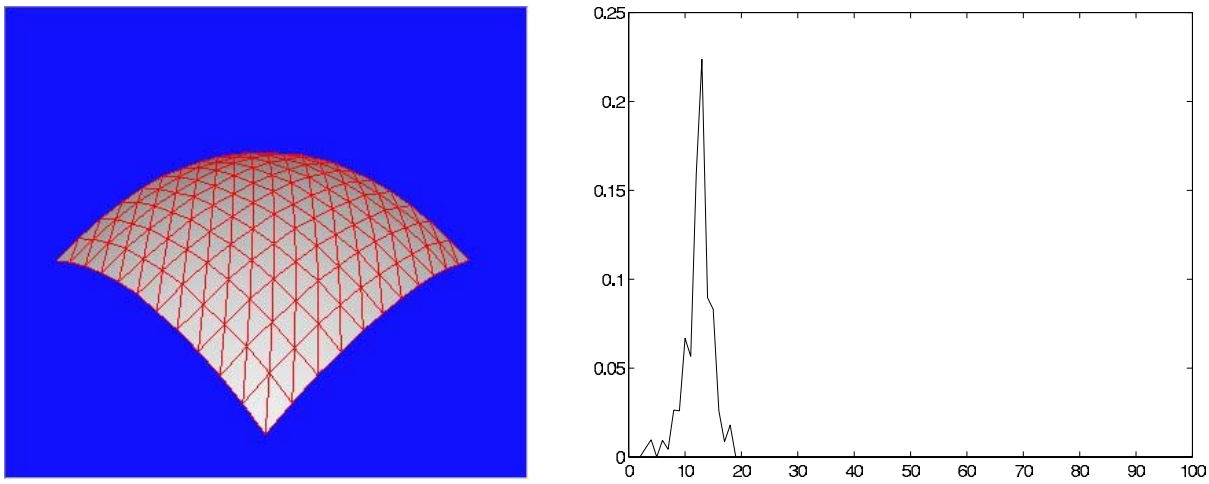
- Les maillages considérés sont maillés de manière suffisamment dense pour que les différentes différences finies considérées approchent raisonnablement les entités infinitésimales associées.
- Les maillages satisfont des contraintes supplémentaires de régularité géométrique (tailles et des rapports d'aspect comparables et adaptés aux courbures locales des surfaces) et topologique (maillages simplexes ou triangulations de topologies équivalentes à la sphère).

Lorsque ces conditions ne sont pas satisfaites, les approximations discrètes par différences finies s'écartent significativement des valeurs correspondantes sur les surfaces continues.

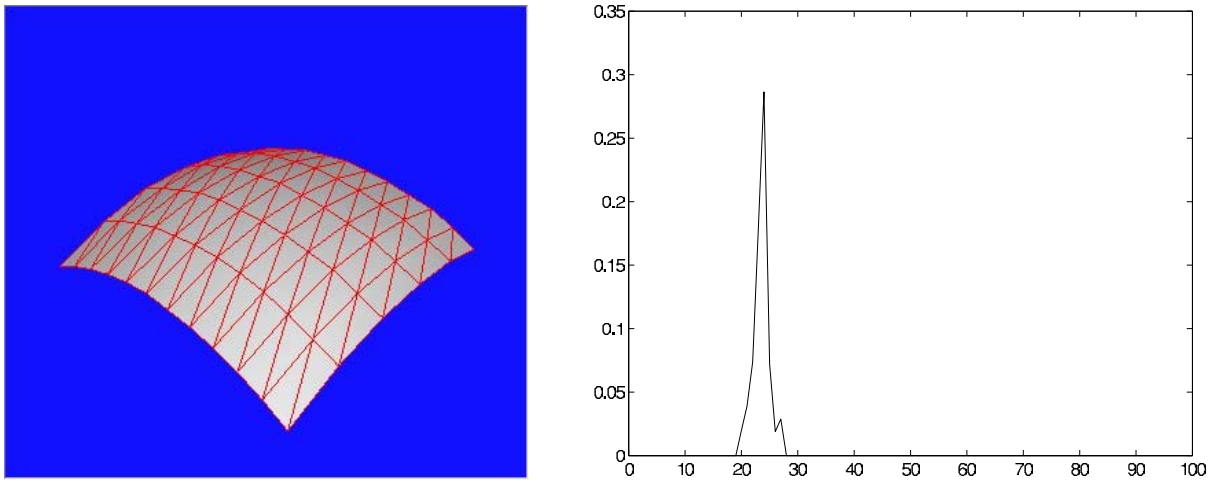
Malheureusement, ces hypothèses sont rarement vérifiées en pratique, lorsque l'on considère des maillages quelconques comme ceux disponibles sur Internet. En outre, ces opérateurs différentiels sont d'une manière générale très sensibles au bruit susceptible d'affecter les valeurs des coordonnées (comme, par exemple, le bruit de quantification dans les maillages ayant fait l'objet d'une procédure de compression). Ces méthodes d'estimation se révèlent également très sensibles aux diverses triangulations du même maillage, comme illustré Figure 3.7. L'algorithme [Taubin95] estime la distribution idéale (Dirac à l'origine) uniquement pour l'échantillonnage de la sphère par un icosaèdre. Les distributions associées aux patchs sphériques définis sur des treillis rectangulaires sont non seulement très éloignées de celle théorique, mais leur aspect varie drastiquement d'une triangulation à l'autre.




a. Icosaèdre.




b. *Patch* sphérique sur treillis carré.




c. *Patch* sphérique sur treillis rectangulaire.

**Figure 3.7.** Différents échantillonnages de la sphère et distributions de l'index de forme associé (avec l'algorithme d'estimation des courbures principales proposé dans [Taubin95]).

En pratique, les maillages ne satisfont jamais les contraintes de régularité géométrique et topologique mentionnées ci-dessus. La solution consiste alors en un remaillage complet de la surface, comme par exemple en utilisant les méthodes proposées dans [Delingette91, Delingette92, Cohen92], afin d'obtenir des représentations conformes à ces différentes contraintes. Ces méthodes s'appuient en général sur l'ajustement d'un maillage déformable, et nécessitent de minimiser itérativement une certaine fonctionnelle d'erreur d'ajustement. Toutefois, ces méthodes requièrent en général des raffinements complémentaires lorsqu'il s'agit de surfaces arbitraires qui peuvent ne pas être homotopes à la sphère (surfaces présentant de multiples composantes connexes ou des trous). De plus, en pratique, la complexité des algorithmes d'optimisation considérés rend leur utilisation prohibitive en temps de calcul.
Comme il est difficilement concevable de remailler tous les modèles disponibles aujourd'hui sur *Internet* afin de les rendre géométriquement exploitables, nous ne retiendrons pas cette première voie méthodologique.
C'est pourquoi nous nous attachons à présent à mettre en perspective de notre problématique de calcul des courbures, les approches paramétriques.

### 3.2.3.2 Approches paramétriques

A l'opposé des méthodes non-paramétriques, le principe des approches paramétriques consiste à associer aux maillages 3D les éléments différentiels estimés à partir d'une approximation du maillage sous forme d'une surface 3D paramétrique régulière.

Quel type de représentation, implicite ou explicite, adopter pour cette surface 3D ?

Dans le cadre de la modélisation paramétrique par fonction implicite, une surface implicite est définie comme l'ensemble des zéros $Z(F_\theta)$ d'une fonction $F_\theta$, définie de $\mathbb{R}^3$ à valeurs dans $\mathbb{R}$, dépendant d'un vecteur de paramètres réels, noté $\theta$, donné par :

$$Z(F_\theta) = \left\{ (x, y, z) \in \mathrm{IR}^3 \mid F_\theta(x, y, z) = 0 \right\}. \tag{3.30}$$

L'ajustement de la surface paramétrique à un ensemble de points $\left\{(x_i, y_i, z_i)\right\}_{i=1}^{N}$ se fait en minimisant, dans l'espace des paramètres $\theta$, une fonctionnelle d'erreur. Celle-ci peut être soit algébrique, s'exprimant sous la forme suivante :

$$\sum_{i=1}^{N} F_\theta^2(x_i, y_i, z_i), \tag{3.31}$$

soit géométrique, et prendre en compte l'information de distance des points à la surface d'ajustement. Une telle approximation est proposée dans [Taubin91]. Elle résulte d'un développement de Taylor au premier ordre de $F_\theta$ et s'exprime par :

$$\sum_{i=1}^{N} \frac{F_\theta^2(x_i, y_i, z_i)}{\left\| \nabla F_\theta^2(x_i, y_i, z_i) \right\|^2}. \tag{3.32}$$

Notons que la solution optimale est dépendante du repère initialement associé au nuage de points.

Soulignons également que, le plus souvent, la minimisation s'effectue sur un sous-ensemble de l'espace paramétrique, ce qui se traduit par l'introduction de contraintes au niveau de la fonctionnelle d'erreur. Sans discuter plus avant du fait que ces contraintes relèvent d'heuristiques, il est important de remarquer qu'elles doivent satisfaire à des propriétés d'invariance au changement de repère, afin que l'invariance géométrique des caractéristiques différentielles, qui est intrinsèque à leur définition, reste garantie.

Un autre inconvénient de la représentation par fonction implicite réside dans l'instabilité des solutions obtenues, de petites variations sur les données pouvant conduire à des solutions totalement différentes. Injecter des propriétés de continuité, directe et inverse, relève d'approches plus ou moins *ad-hoc* et plus ou moins sophistiquées, comme par exemple dans [Blane00], où la procédure consiste à ajuster une surface implicite dans $\mathbb{R}^k$ conditionnellement à une famille discrète de courbes de niveaux d'une surface explicite de $\mathbb{R}^{k+1}$. Bien que mathématiquement fondée, cette solution est difficilement transposable au cas des surfaces 3D maillées dont les coordonnées des sommets ne sont pas définies sur un treillis régulier.

Enfin, considérer des fonctionnelles géométriques conduit à résoudre un système d'équations fortement non-linéaires pour lequel il est nécessaire de recourir à des méthodes itératives de minimisation, comme par exemple celle de Levenberg-Marquardt [NRC92]. Se posent alors les problèmes bien connus de convergence et de coût de calcul.

L'analyse de ces différents inconvénients des méthodes d'ajustement de surfaces paramétriques représentées par des fonctions implicites, nous a conduit à considérer un deuxième type de représentation paramétrique des surfaces 3D, à l'aide de fonctions explicites.

Une fonction explicite définie à partir d'une paramétrisation de Monge dans un repère local ($x$, $y$, $z$) à la surface e'expriment sous la forme :

$$S_a = (x, y, f_a(x,y)), \tag{3.33}$$

où $a$ désigne le vecteur de paramètres.
Parmi les différentes fonctions $f_a$ possibles (polynomiales, splines, ...) [Flynn89, McIvor97, Lengagne96, Greiner96], nous avons choisi, pour des raisons de simplicité analytique et de régularité différentielle, une fonction quadratique.

La fonctionnelle d'erreur de l'ajustement de cette surface au maillage considéré est donnée par l'équation générique ci-dessous :

$$\mathrm{E} = \sum_{i=1}^{N} w_i \left(z_i - f_a(x_i, y_i)\right)^2 , \tag{3.34}$$

où $w_i$ représente le poids associé au point $(x_i, y_i, z_i)$.
Là encore, les solutions dépendent du repère local donné. Mais contrairement aux méthodes d'ajustement par fonction implicite, l'estimation des paramètres ne nécessite pas ici d'imposer des contraintes supplémentaires et peut être réalisée par une simple régression linéaire.

De cette analyse des méthodes d'ajustement, il ressort que malgré leur généricité, les fonctions implicites conduisent à résoudre des systèmes à la fois complexes et mal conditionnés. L'approche par fonction explicite quadratique avec paramétrisation de Monge offre en revanche toutes les propriétés d'invariance souhaitées en même temps que des schémas d'estimation des courbures simples et robustes que nous décrivons dans le paragraphe suivant.

#### 3.2.3.3 Estimation des courbures par approximation quadratique

Les développements ci-dessous s'appliquent à des maillages localement approchés par des surfaces quadratiques avec paramétrisation de Monge.
Les courbures principales sont associées à chaque facette du maillage. Le vecteur normal moyen associé à la facette $f_i$, noté $\tilde{N}_{f_i}$, est défini comme la moyenne des vecteurs normaux à toutes les facettes 0-adjacentes à $f_i$ (*i.e.* ayant au moins un sommet en commun avec $f_i$). L'expression de $\tilde{N}_{f_i}$ est donnée par :

$$\tilde{N}_{f_i} = \frac{\sum_{f_k \in F_0\{f_i\}} w_k N_{f_k}}{\left\| \sum_{f_k \in F_0\{f_i\}} w_k N_{f_k} \right\|}, \tag{3.35}$$

où

- $F_0\{f_i\}$ désigne l'ensemble des facettes 0-adjacentes à $f_i$,
- $N_{f_k}$ est le vecteur normal à la facette $f_k$ et $w_k$ le poids associé à la facette $f_k$,
- $\|\cdot\|$ est la norme $L_2$.

Comme les tailles des facettes du maillage peuvent être très différentes, les poids $w_k$ sont choisis égaux aux aires des facettes.

Pour chaque facette $f_i$, un repère cartésien local est construit. Son origine est placée au centre de gravité de $f_i$ et l'axe des $z$ est défini selon la direction du vecteur normal moyen. Dans ce repère, considérons une surface polynomiale du second degré $S_a = (x, y, f_a(x,y))$, avec

$$f_a(x,y) = a_0 x^2 + a_1 y^2 + a_2 xy + a_3 x + a_4 y + a_5 , \tag{3.36}$$

où les coefficients $a_i$ prennent des valeurs réelles.

En notant : $a = (a_0\, a_1\, a_2\, a_3\, a_4\, a_5)^\tau$ et $b(x,y) = (x^2\ y^2\ xy\ x\ y\ 1)^\tau$, l'équation (3.36) s'écrit :

$$f_a(x,y) = a^\tau b(x,y)\,, \tag{3.37}$$

Soit $\{(x_i,y_i,z_i)\}_{i=1}^N$ l'ensemble des centres de gravité de la facette considérée et de ses facettes 0-adjacentes (leurs coordonnées étant exprimées dans le repère local). Le vecteur de paramètres $\hat{a}$ optimal au sens de la minimisation de l'erreur quadratique moyenne pondérée $\mathrm{E} = \sum_{i=1}^{N} w_i (z_i - f_a(x_i,y_i))^2$ est donné par l'équation suivante :

$$\hat{a} = \left( \sum_{i=1}^{N} w_i b(x_i,y_i) b^\tau(x_i,y_i) \right)^{-1} \left( \sum_{i=1}^{N} w_i z_i b(x_i,y_i) \right) = A^{-1} v\,, \tag{3.38}$$

où

$$A = \left( \sum_{i=1}^{N} w_i b(x_i,y_i) b^\tau(x_i,y_i) \right) \text{ et } v = \left( \sum_{i=1}^{N} w_i z_i b(x_i,y_i) \right). \tag{3.39}$$

En pratique, l'inversion de la matrice *A* est effectuée par la méthode par pseudo-inverse afin de garantir la stabilité de la solution.

Etudions à présent le comportement de l'ajustement quadratique $\hat{a}$ en terme d'invariance par rapport aux transformations de similarité.

Remarquons immédiatement que la représentation des centres de gravité dans le repère local rend *de facto* l'approximation $\hat{a}$ invariante aux transformations euclidiennes. La proposition suivante explicite le comportement du vecteur de paramètres par rapport aux homothéties.

**Proposition 3.3**

Soient $\hat{a}$ et $\hat{a}_\alpha$, les vecteurs de paramètres respectivement associés à l'ensemble des points $P = \{(x_i,y_i,z_i)\}_{i=1}^N$ et $P_\alpha = \{(\alpha x_i, \alpha y_i, \alpha z_i)\}_{i=1}^N$, où $\alpha$ est un facteur d'homothétie non nul. La relation entre $\hat{a}$ et $\hat{a}_\alpha$ est donnée par :

$$\hat{a}_\alpha = H_\alpha \hat{a}\,, \tag{3.40}$$

où

$$H_\alpha = \begin{pmatrix} \alpha^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha \end{pmatrix} \overset{not}{=} diag(\alpha^{-1}, \alpha^{-1}, \alpha^{-1}, 1, 1, \alpha), \tag{3.41}$$

*Démonstration*

Soient $A_\alpha$ et $v_\alpha$ les valeurs relatives à $P_\alpha$, exprimées par :

$$A_\alpha = \left( \sum_{i=1}^{N} w_i b(\alpha x_i, \alpha y_i) b^\tau(\alpha x_i, \alpha y_i) \right), \; v_\alpha = \left( \sum_{i=1}^{N} w_i \, \alpha z_i \, b(\alpha x_i, \alpha y_i) \right) \tag{3.42}$$

Remarquons que $b(\alpha x_i, \alpha y_i)$ s'exprime comme le produit de $b(x_i, y_i)$ par une matrice diagonale :

$$b(\alpha x_i, \alpha y_i) = \begin{pmatrix} \alpha^2 x_i^2 \\ \alpha^2 y_i^2 \\ \alpha^2 x_i y_i \\ \alpha x_i \\ \alpha y_i \\ 1 \end{pmatrix} = diag(\alpha^2, \alpha^2, \alpha^2, \alpha, \alpha, 1) \begin{pmatrix} x_i^2 \\ y_i^2 \\ x_i y_i \\ x_i \\ y_i \\ 1 \end{pmatrix} = diag(\alpha^2, \alpha^2, \alpha^2, \alpha, \alpha, 1) \, b(x_i, y_i), \tag{3.43}$$

en notant $D_\alpha = diag(\alpha^2, \alpha^2, \alpha^2, \alpha, \alpha, 1)$, on a $b(\alpha x_i, \alpha y_i) = D_\alpha b(x_i, y_i)$.

Par substitution dans $A_\alpha$ et par commutativité du produit de deux matrices lorsque l'une d'entre elles est diagonale, on obtient l'expression ci-dessous :

$$A_\alpha = \sum_{i=1}^{N} w_i D_\alpha b(x_i, y_i) b^\tau(x_i, y_i) D_\alpha = D_\alpha^2 \sum_{i=1}^{N} w_i b(x_i, y_i) b^\tau(x_i, y_i) = D_\alpha^2 A, \tag{3.44}$$

De manière similaire, on montre que

$$v_\alpha = \alpha \, D_\alpha v. \tag{3.45}$$

Le vecteur de paramètres $\hat{a}_\alpha$ s'exprime alors sous la forme :

$$\hat{a}_\alpha = (D_\alpha^{-2} A^{-1})(\alpha \, D_\alpha v) = (\alpha \, D_\alpha^{-1})(A^{-1} v) = H_\alpha \hat{a}, \tag{3.46}$$

où

$$H_\alpha = \alpha D_\alpha^{-1} = diag(\alpha^{-1}, \alpha^{-1}, \alpha^{-1}, 1, 1, \alpha), \tag{3.47}$$

ce qui établit la relation de la Proposition 3.3.

Disposant de l'approximation polynomiale, les deux formes différentielles, calculées à l'origine (0, 0) de

la paramétrisation de Monge, s'expriment comme suit :

$$I = \begin{pmatrix} 1+a_3^2 & a_3 a_4 \\ a_3 a_4 & 1+a_4^2 \end{pmatrix},\ II = \frac{1}{\sqrt{1+a_3^2+a_4^2}} \begin{pmatrix} a_0 & a_2 \\ a_2 & a_1 \end{pmatrix}. \tag{3.48}$$

Il en résulte que la matrice de Weingarten $A$ dépend du facteur $\alpha$ uniquement par l'intermédiaire des termes quadratiques $a_0$, $a_1$ et $a_2$ intervenant dans l'expression de la deuxième forme différentielle. En conséquence, les courbures principales (valeurs propres de $A$) sont inversement proportionnelles à $\alpha$. On retrouve ainsi la propriété bien connue dans le cas des surfaces continues. Il s'ensuit que le rapport des courbures principales est indépendant du facteur d'échelle $\alpha$.

Remarques :

1. La propriété que nous venons d'énoncer est également valable pour toute approximation polygonale de degré $n \geq 2$. Cela résulte du fait que les termes de degré supérieur à 2 sont des monômes en $x$ et $y$ qui s'annulent à l'origine (0, 0) de la paramétrisation. En outre, les expressions des deux formes différentielles restent identiques à celles de l'équation (3.48).
2. Pour accentuer l'effet de lissage, il est possible, pour estimer les paramètres de la surface, de considérer à la place du voisinage $F_0\{f_i\}$, le voisinage $F_n\{f_i\}$, définissant les facettes n-adjacentes ($n > 0$).

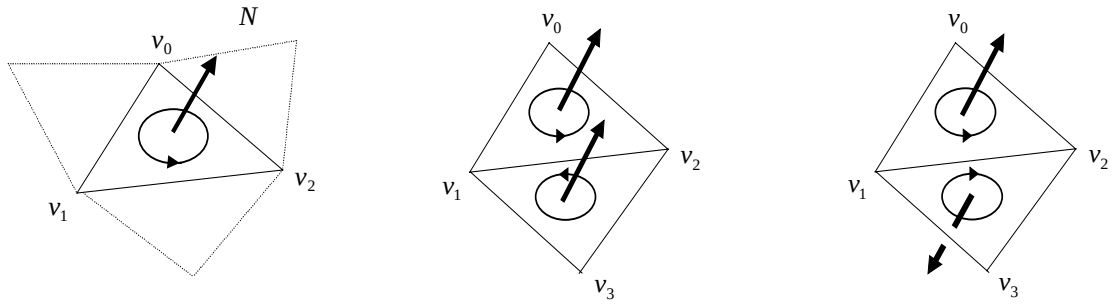
En résumé, le schéma d'estimation décrit ci-dessus offre une solution simple pour estimer les courbures principales. Néanmoins, pour qu'il soit applicable, les maillages considérés doivent respecter certaines hypothèses d'orientation et de régularité, que nous précisons par la suite.

### 3.2.3.4 Orientation et régularité des maillages 3D

Introduisons tout d'abord quelques définitions pour préciser mathématiquement le contexte. Deux sommets distincts successifs (modulo une permutation cyclique) dans une séquence de sommets définissant une facette forment une *arête*. Une arête appartenant à une unique facette constitue une *arête de bord*. Les *facettes de bord* sont définies comme des facettes contenant au moins une arête de bord. Les arêtes qui ne sont pas des arêtes de bord sont contenues dans au moins deux facettes différentes et sont appelées *arêtes intérieures*. Deux facettes sont *A-voisines* si elles partagent au moins une arête commune. Un sous-ensemble de facettes du maillage est dit *connexe*, si pour deux facettes distinctes quelconques, il existe un chemin de facettes successives A-voisines les reliant. Dans la suite, une *composante connexe* du maillage désignera un sous-ensemble connexe maximal (au sens du nombre de facettes) du maillage.

Par analogie avec l'interprétation du concept d'orientation rappelé au Paragraphe 3.2.1, chaque facette du maillage est orientée selon le vecteur normal qui lui est orthogonal et dont le sens est donné par l'ordre de parcours des sommets (Figure 3.8.a). Un maillage est dit *orientable* si deux facettes A-voisines quelconques ont des orientations cohérentes (Figure 3.8.b), *i.e.* elles sont parcourues dans des sens opposés.

La Figure 3.8.c montre un exemple de maillage non-orientable : l'arête ($v_1$, $v_2$) commune aux deux *facettes A-voisines* est parcourue dans le même sens.
Remarquons qu'une arête d'un maillage orientable appartient à au plus deux facettes.




a. Orientation d'une facette. b. Maillage orientable. c. Maillage non-orientable.

**Figure 3.8.** Principe d'orientation d'un maillage.

Pour calculer l'index de forme et les courbures principales du maillage, il est donc nécessaire d'imposer à ce dernier d'être orientable.
Cette contrainte n'est pas suffisante puisque le calcul des dérivées ne fait sens que si la surface maillée est suffisamment "lisse", i.*e.* au moins $C^2$- différentiable. Pour garantir cette régularité différentielle, nous proposons d'appliquer de façon systématique à chaque maillage, l'algorithme de subdivision de Loop [Loop87].
Il consiste en les deux étapes suivantes : (1) re-échantillonnage du maillage par insertion de sommets aux milieu de chaque arête (Figure 3.9) ; (2) filtrage passe-bas des coordonnées des sommets.

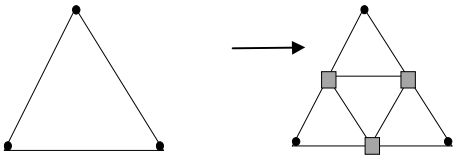


**Figure 3.9.** Principe de ré-échantillonnage du maillage :
un sommet est inséré au milieu de chaque arête du maillage initial .

Soient $v_0$ un sommet quelconque du maillage maillage ré-échantillonné et $\{v_k\}_{k=1}^{N}$ ses sommets voisins. En notant $\{\alpha_k\}_{k=0}^{N}$ les coefficients de pondération associés à chacun des sommets, la nouvelle position, $v_0^{new}$, de $v_0$ s'exprime comme une combinaison linéaire des sommets voisins (Figure 3.10) :

$$v_0^{new} = \frac{\alpha_0 v_0 + \sum_{k=1}^{N} \alpha_k v_k}{\alpha_0 + \sum_{k=1}^{N} \alpha_k}, \tag{3.49}$$

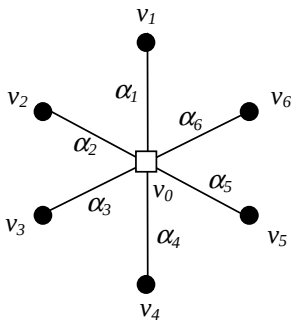


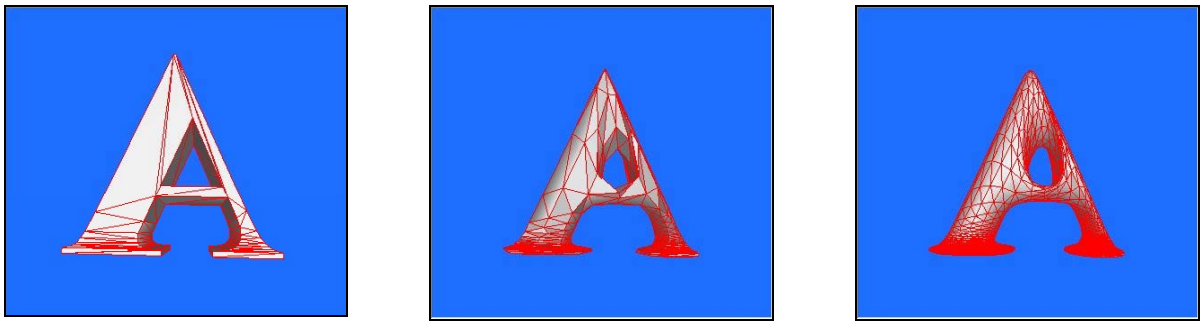

**Figure 3.10.** L'ensemble des sommets voisins de $v_0$ et les poids associés.

Pour une surface de subdivision de Loop, les poids prennent les valeurs suivantes :

- $\alpha_0 = \dfrac{N\beta(N)}{1-\beta(N)}$ , avec $\beta(N) = 2\left(\dfrac{3}{8} + \dfrac{\cos(\frac{2\pi}{N})}{4}\right)^2 - \dfrac{1}{4}$, (3.50)
- $\forall\, k \in \{1,2,\ldots,N\},\ \alpha_k = 1$.

Selon le choix des poids $\{\alpha_k\}_{k=0}^{N}$, les surfaces limites (*i.e.* obtenues en itérant à l'infini l'algorithme) satisfont des propriétés différentes en terme de dérivabilité [Catmull74, Taubin95fair, Lounsberry94, DeRose98, Desbrun99]. On peut montrer que dans le cas de la subdivision de Loop, les surfaces limites résultantes sont $C^2$- continues [Lounsberry94], ce qui est la propriété minimale que les maillages doivent satisfaire dans notre cas. Le lissage est illustré Figure 3.11.

Notons en outre que l'application d'un tel algorithme de subdivision présente l'avantage de réduire l'aire des surfaces de bord par rapport à l'aire totale du maillage, où l'estimation de l'index de forme est moins fiable en raison du nombre réduit de facettes 0-adjacentes à une facette de bord.



a. Maillage initial. b. Maillage après une subdivision. c. Maillage après deux subdivisions.

**Figure 3.11.** Exemple d'application de la subdivision de Loop :
notons l'effet progressivement régularisant obtenu.

Désormais, nous ne considérons que des maillages à la fois orientables et $C^2$- régularisés, les courbures principales pouvant alors être estimées selon le schéma précédemment exposé.

## 3.2.4 Le descripteur MPEG-7 par spectre de forme 3D

Pour que le SF3D, tel que présenté ci-dessus, devienne un descripteur normalisé MPEG-7, il est nécessaire de spécifier la syntaxe et la sémantique de chaque élément constitutif du descripteur.
Présentons donc le descripteur MPEG-7 par spectre de forme 3D.

### 3.2.4.1 Définition et interprétation

Le SF3D d'un maillage est défini comme l'histogramme de l'index de forme. Considérons une subdivision uniforme $\{\Delta_k\}_{k=1}^{N_{bins}}$ de l'intervalle [0, 1] en un nombre $N_{bins}$ d'intervalles quantifiant uniformément le domaine des valeurs de l'index de forme (l'intervalle [0, 1]), avec

$$\forall\ k \in \{1,2,\ldots,N_{bins}-1\},\quad \Delta_k = \left[\frac{k-1}{N_{bins}}, \frac{k}{N_{bins}}\right), \text{ et}$$

$$\Delta_{N_{bins}} = \left[\frac{N_{bins}-1}{N_{bins}}, 1\right].$$

Le SF3D est représenté comme un vecteur à $N_{bins}$ composantes, la composante $k$ cumulant l'aire relative (par rapport à l'aire totale du maillage) de toutes les facettes ayant un index de forme appartenant à $\Delta_k$.
Deux éléments supplémentaires sont ajoutés au descripteur. Le premier, noté *SurfacesPlanes,* quantifie l'aire relative des surfaces planes (pour lesquelles l'index de forme n'est pas défini) ; le second, *SurfacesSingulières,* celle des facettes de bord ; le nombre réduit de facettes 0-adjacentes ne permettant pas d'effectuer un ajustement polynomial fiable.
Pour décider si une surface est localement plane ou non, nous calculons tout d'abord la norme $L_2$ quadratique du vecteur de courbures , notée $k_a^2$ et exprimée par :

$$k_a^2 = k_1^2 + k_2^2 . \tag{3.51}$$

Le degré de planéité est ensuite défini par le critère $C$ comme suit :

$$C = Aire(F_0\{f_i\}) \cdot k_a^2 . \tag{3.52}$$

Si $C$ est inférieur à un seuil pre-déterminé, alors la facette $f_i$ est dite plane.
Notons que cette définition du degré de planéité assure son invariance par rapport aux homothéties.
La Figure 3.12 présente, pour 3 maillages de complexité croissante, les SF3D associés.

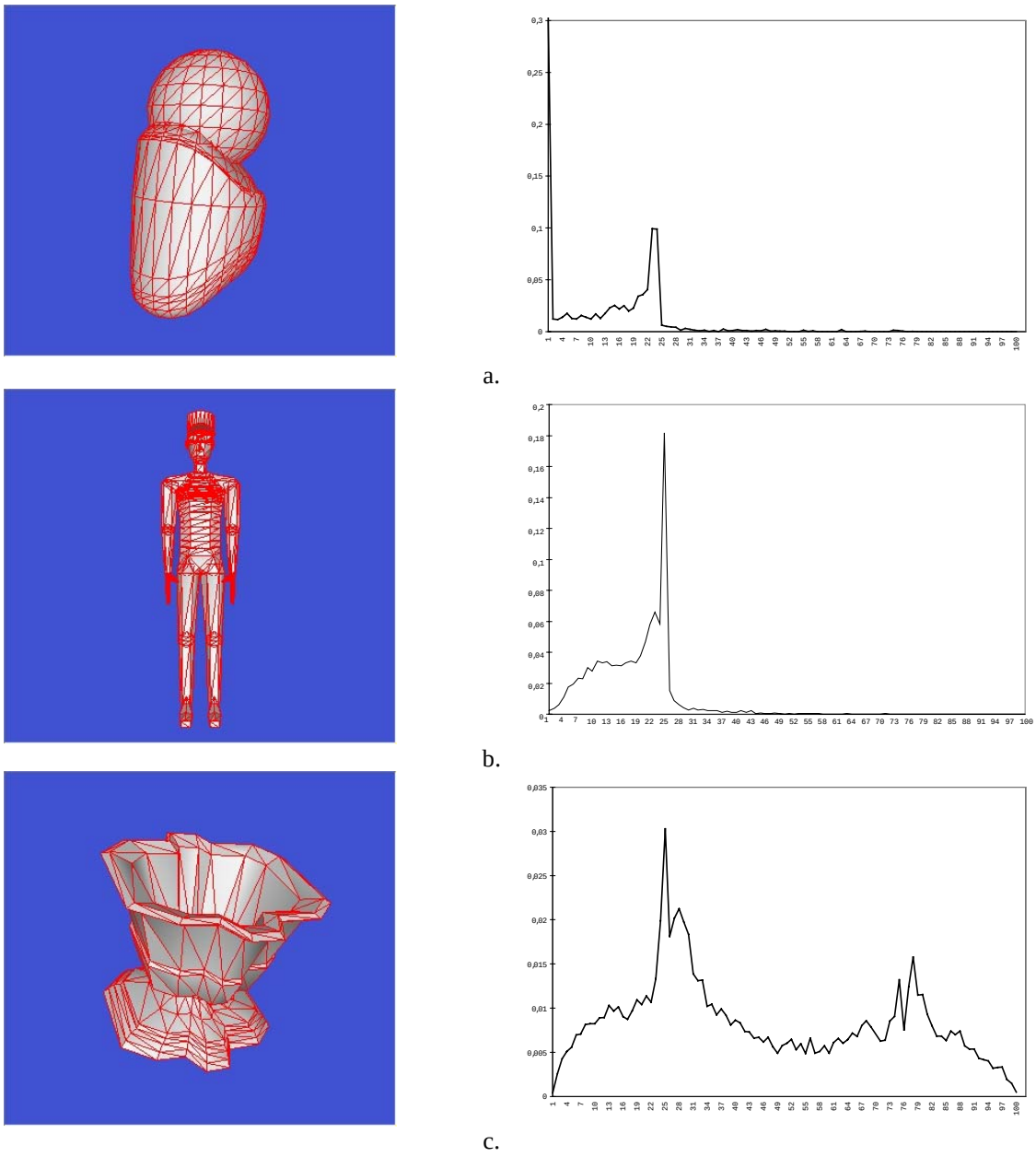



**Figure 3.12.** Maillages 3D et les SF3D associés.

Le maillage illustré Figure 3.12.a est principalement composé d'une surface sphérique et d'une autre cylindrique. En conséquence, le spectre est distribué selon deux maxima prédominants, correspondant aux valeurs de l'index de forme 0 (caractéristique d'une sphère) et 0.25 (caractéristique d'un cylindre). L'humanoïde maillé illustré Figure 3.12.b est un assemblage d'éléments cylindriques. Le SF3D associé représente une distribution mono-modale centrée sur la valeur 0.25. Enfin, le modèle de coupe (Figure 3.12.c) est un exemple d'objet creux présentant donc deux faces, une intérieure et l'autre extérieure. Les éléments convexes et concaves sont quasi également représentés dans le spectre qui comporte deux pics bien distincts, autour des valeurs d'index de forme de 0.25 et 0.75, respectivement.

La syntaxe *XML Schema* du SF3D, telle que retenue dans le standard au stade FDIS, est présentée Figure **3.13**.

```
<complexType name="Shape3DType" final="#all">
   <complexContent>
      <extension base="mpeg7:VisualDType">
         <sequence>
            <element name="Spectrum">
               <simpleType>
                  <restriction>
                     <simpleType>
                        <list itemType="mpeg7:unsigned12"/>
                     </simpleType>
                     <maxLength value="255"/>
                  </restriction>
               </simpleType>
            </element>
            <element name="PlanarSurfaces" type="mpeg7:unsigned12"/>
            <element name="SingularSurfaces" type="mpeg7:unsigned12"/>
         </sequence>
         <attribute name="bitsPerBin" type="mpeg7:unsigned4"
          use="optional" default="12"/>
      </extension>
   </complexContent>
</complexType>
```

**Figure 3.13.** Représentation MPEG-7 du descripteur SF3D.

Enfin, concluons notre discussion sur les aspects normatifs relatifs au SF3D par un bref historique de sa soumission au WG11, en soulignant les principales contributions que nous avons apportées pendant le processus de standardisation.

#### 3.2.4.2 Contexte normatif MPEG-7

Le processus de normalisation d'un descripteur de forme 3D a été initié en juillet 1999, quand la proposition présentée par IBM-Japon et NSF (National Science Foundation) dans [Murao-Iso99.07] a suscité un écho favorable au sein du groupe Vidéo de MPEG-7, qui a approuvé le lancement d'un *Core Experiment* (CE) pour l'évaluation des descripteurs de forme 3D [SM-CE99.07]. Le CE a été alors rejoint par des représentants de Mitsubishi-UK, HHI (Heinrich Hertz Institut), Phillips et nous-mêmes (INT).

Sur proposition de l'INT, la base de données de 50 modèles intialement constituée par IBM-Japon a été enrichie par les 300 modèles disponibles au sein du WG11 et précédemment utilisés dans le cadre des expérimentations MPEG-4 SNHC portant sur l'évaluation des technologies de compression de maillages 3D.

Lors de la réunion internationale du WG11 d'octobre 1999, nous avons soumis un nouveau descripteur, le spectre de forme 3D [Zaharia-Iso99.10s]. Les résultats alors présentés, bien que préliminaires, montraient nettement la compétitivité du SF3D par rapport aux descripteurs initialement proposés dans [Murao-Iso99.02] et [Paquet-Iso99]. Le groupe Vidéo a décidé alors d'encourager les participants du CE à poursuivre les expérimentations par des vérifications croisées à base d'implantations indépendantes et de présenter à la prochaine réunion une étude comparée des différents résultats.

INT et IBM-Japon sont restés les seuls partenaires actifs du CE devant évaluer les technologies proposées. L'implantation des deux technologies, le SF3D et le DMS (descripteur de moments sphériques), a été conduite à son terme par les deux partenaires. Les résultats demandés par le groupe Vidéo présentés en réunion MPEG en Décembre 1999 [Zaharia-Iso99.12s] ont prouvé la supériorité des performances du SF3D qui devint alors la seule technologie de forme 3D considérée dans MPEG-7.

Toutefois, le groupe Vidéo a demandé des investigations supplémentaires dans le cadre d'une base de maillages 3D catégorisée et élargie à au moins 1000 modèles. S'est alors posé le problème de trouver une telle base susceptible d'être mise à disposition de tous les participants MPEG-7 afin de garantir la transparence des évaluations. Des négociations ont été engagées par le *Chair* des CEs sur les descripteurs de forme avec la Société *Viewpoint*, propriétaire d'une large collection de modèles 3D, afin d'utiliser une sous-partie de son corpus pour les expérimentations MPEG-7. Cette négociation a abouti en début d'année 2000 à une proposition dans laquelle les conditions de licence très restrictives et le prix exorbitant des produits *Viewpoint*, en totale rupture avec l'esprit MPEG et les objectifs purement scientifiques des expérimentations, ont conduit le groupe à décliner cette offre et à rechercher d'autres solutions.
Nous avons proposé [Zaharia-Iso00.03a] lors de la réunion WG11 de Mars 2000 d'exploiter un ensemble d'environ 1000 modèles de la collection *3D Cafe* [3DCafe], disponible sur *Internet* sous licence de type "*Royalty Free*". INT a converti tous ces modèles, représentés initialement sous des formats aussi divers que 3DS, DXF, LWO, CBO, au format VRML 2.0, pour le confort des participants au CE. Parallèlement, le SF3D avait fait à nouveau l'objet d'une vérification croisée [Zaharia-Iso00.03b, Hoogvorst-Iso00.03], avec une nouvelle implantation indépendante et les résultats obtenus confirmaient ceux de Décembre 1999.

Pour des raisons techniques au niveau du WG11 une réunion supplémentaire a été ajoutée au calendrier ISO, en juin 2000. Nous en avons profité pour soumettre deux études génériques, l'une sur la manière de définir des catégories pertinentes [Zaharia-Iso00.06a] et la seconde sur les problèmes de représentations topologiques multiples associées aux données maillées [Zaharia-Iso00.06b] qui ont été toutes deux jugées pertinentes et prises en compte par le groupe Vidéo. Une retombée immédiate a été la catégorisation de la base de modèles *3D Cafe* selon la proposition de l'INT.
Ces avancées méthodologiques une fois acquises ainsi que la mise en place précise du protocole d'évaluation quantitative du descripteur ont permis à l'INT, pendant la réunion MPEG de juillet 2000, de présenter les résultats objectifs et complets sur l'ensemble des catégories [Zaharia-Iso00.07a, Zaharia-Iso00.07b]. Cela a conduit à l'adoption du SF3D dans le standard MPEG-7 (à cette époque, à l'état de *Working Draft*).

L'INT a ensuite développé et intégré, conformément aux procédures de normalisation MPEG, le SF3D dans le XM MPEG-7 (Août 2000) et assure depuis la maintenance du descripteur, à travers des mises à jour logicielles pour suivre les évolutions naturelles du standard (comme par exemple l'intégration du support DDL).

Le SF3D est le seul descripteur de forme 3D adopté dans le futur standard MPEG-7. Par construction, il

est invariant aux transformations euclidiennes et aux homothéties. Etudions à présent son comportement vis-à-vis des représentations topologiques multiples.

### 3.2.5 De l'invariance topologique du SF3D

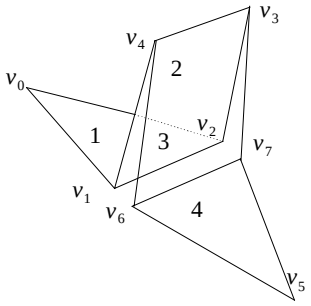
Les modèles 3D sont en général maillés en vue d'applications de type graphique informatique qui privilégient les critères visuels au détriment d'une certaine régularité topologique des représentations créées. Dans ce contexte, les maillages réalisés peuvent se présenter sous forme de multiples descriptions maillées, topologiquement complètement différentes et inclure des polygones dégénérés ou dupliqués. Analysons l'influence de ces divers problèmes de représentation topologique sur le comportement du SF3D.

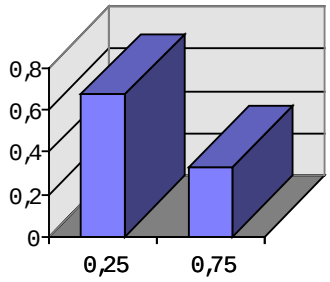
Considérons tout d'abord les quatre maillages orientés, illustrés Figure 3.14. Ils représentent le même objet, formé de 4 facettes (définies entre accolades et numérotées de 1 à 4) identiquement orientées. Pour un *Viewer* VRML, ces quatre maillages sont visuellement identiques. Cependant, les surfaces géométriques associées sont complètement différentes. Ainsi, le maillage de la Figure 3.14.a correspond-il à une représentation minimale au sens où il est constitué d'une seule composante connexe. Pour le maillage de la Figure 3.14.b, chaque facette représente une composante connexe. Les maillages des Figure 3.14.c et d comprennent chacun deux composantes connexes, formées des facettes (1, 2 et 3) et (4) et des facettes (1, 2) et (3, 4), respectivement.
L'analyse géométrique de ces surfaces en terme d'index de forme, ou, plus généralement en termes d'attributs différentiels de surface, conduit à des résultats bien différents, comme illustré Figure 3.14. Cela démontre à l'évidence la forte sensibilité du SF3D à ces aspects de représentations topologiques multiples.

Ajoutons que la présence d'éventuels polygones dupliqués ou dégénérés (définis comme polygones d'aire nulle) affecte aussi bien les caractéristiques globales (comme le centre de gravité ou les moments statistiques) que locales de la surface. Dans ce cas, le SF3D est également modifié en fonction de cette dégénérescence ou multiplicité arbitraire.

De cette analyse, il ressort que pour conférer une certaine robustesse au SF3D vis-à-vis des représentations topologiques, il est nécessaire de l'appliquer à des maillages 3D canoniques, i.e. ayant une représentation topologique minimale et régulière, exprimée en terme de surfaces orientables, composée d'un nombre minimal de composantes connexes et sans polygone multiple ou dégénéré.

a.

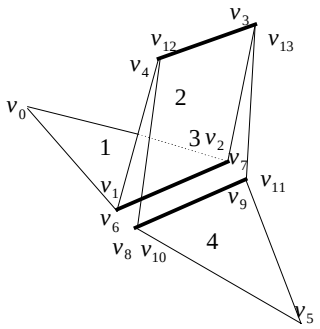


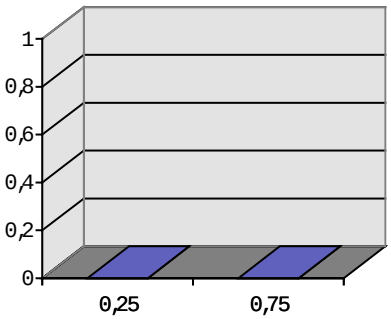




1={v0, v1, v2}, 2={ v1, v4, v3, v2},
3={ v6, v7, v3, v4}, 4={v7, v6, v5}

SF3D(a)

b.

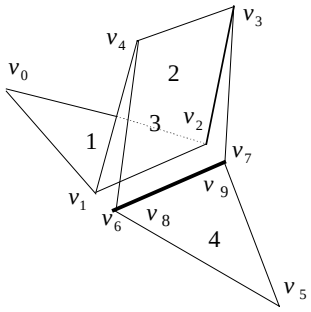


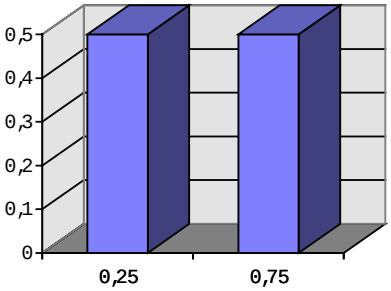




1={v0, v1, v2}, 2={ v6, v4, v3, v7},
3={ v8, v9, v13, v12}, 4={v11, v10, v5}

SF3D(b)

c.

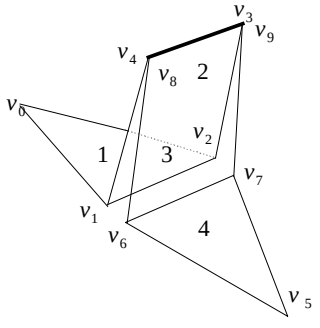


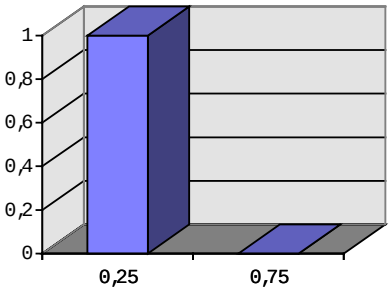




1={v0, v1, v2}, 2={ v2, v1, v4, v3},
3={ v6, v7, v3, v4}, 4={v9, v8, v5}

SF3D(c)

d.




1={v0, v1, v2}, 2={ v1, v4, v3, v2},
3={ v6, v7, v9, v8}, 4={v7, v6, v5}

SF3D(d)

**Figure 3.14.** Quatre représentations topologiques différentes d'un même objet (les arêtes en gras délimitent les composantes connexes) et les SF3D associés.

Pour cela, un filtrage des modèles 3D [Gueziec98, Hoogvorst00.06] a été appliqué. D'autres approches permettant de "réparer" de tels maillages 3D sont proposées dans [Butlin96, Barequet97, Murali97]. Néanmoins, notre expérience montre qu'il n'existe pas de méthode générale permettant de résoudre l'ensemble de ces problèmes.
Pour cette raison, il est pertinent de disposer d'un DF s'affranchissant de l'information de connexité. C'est pour répondre à cet objectif que nous introduisons un DF fondé sur la transformée de Hough 3D.

# 3.3 Description de forme par transformée de Hough 3D

La transformée de Hough [Hough62] a été largement utilisée en analyse d'images, pour des applications aussi diverses que la détection de primitives ou la reconnaissance d'objets 2D ou 3D (Paragraphe 3.1.3). Rappelons-en tout d'abord le principe.

## 3.3.1 La transformée de Hough 3D

La transformée de Hough 3D (TH 3D) est fondée sur un principe d'accumulation des points sur des plans du $\mathbb{R}^3$. Commençons donc par présenter la définition de base et le principe de construction algorithmique.

### 3.3.1.1 Définition et construction algorithmique

Soit $E$ un ensemble fini de points dans $\mathbb{R}^3$, dont les coordonnées sont précisées dans un repère cartésien $(O, x, y, z)$.
Un plan $\Pi$ de $\mathbb{R}^3$ est caractérisé de manière unique par un triplet $(s, \theta, \varphi)$, où

- $s \geq 0$ représente la distance de l'origine du repère au plan $\Pi$,
- $\theta \in [0, 2\pi)$ et $\varphi \in [-\pi/2, \pi/2)$ désignent les deux angles associés à la représentation en coordonnées sphériques du vecteur unitaire $\boldsymbol{n}=(n_x, n_y, n_z)^t$ normal au plan (Figure 3.15).

Les trois composantes de $\boldsymbol{n}$ s'expriment en fonction de $\theta$ et $\varphi$ comme suit :

$$n_x = \cos(\varphi)\cos(\theta),\ n_y = \cos(\varphi)\sin(\theta),\ n_z = \sin(\varphi). \tag{3.53}$$

La distance de l'origine du repère au plan passant par $p$ et de vecteur $\boldsymbol{n}$ s'écrit : $s = \left| n_x x + n_y y + n_z z \right|$.

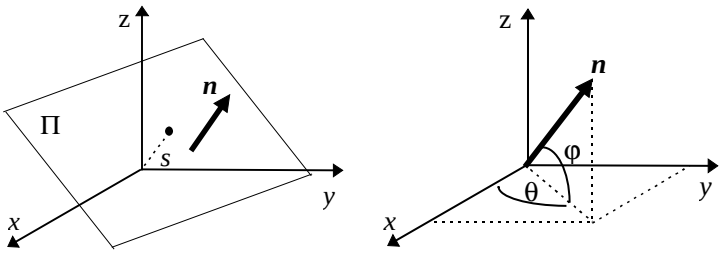



**Figure 3.15.** Paramétrisation des plans de $\mathbb{R}^3$.

En échantillonnant uniformément chaque axe de l'espace des paramètres $(s, \theta, \varphi)$, on obtient

respectivement $N_s$, $N_\theta$ et $N_\varphi$ éléments regroupés dans les ensembles suivants :

- $\Xi = \{ s_k = k\Delta_s \}$,
  où $\Delta_s = S_{\max} / N_s$, $k \in \{0,1,\ldots,N_s - 1\}$ et $S_{max}$ est une valeur suffisamment grande définissant la taille du maillage,
- $\Theta = \{ \theta_k = (k + 0.5)\Delta_\theta \}$,
  où $\Delta_\theta = 2\pi / N_\theta$, $k \in \{0,1,\ldots,N_\theta - 1\}$ et
- $\Psi = \{ \varphi_k = (k - N_\varphi / 2 + 0.5)\Delta_\varphi \}$,
  où $\Delta_\varphi = \pi / N_\varphi$, $k \in \{0,1,\ldots,N_\varphi - 1\}$.

Les ensembles $\Theta$ et $\Psi$ définissent en effet les prototypes de quantification de l'espace des orientations. Représentons ces orientations comme des vecteurs sur la sphère unité. L'échantillonnage uniforme, réalisé de manière indépendante selon chaque paramètre $\theta$ et $\varphi$, conduit alors à une partition de la sphère unité en "parallèles" et "méridiens", comme illustré Figure 3.16.

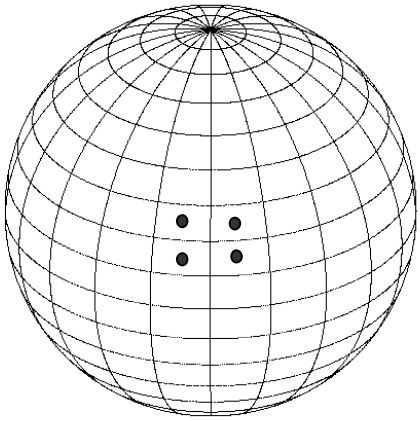


**Figure 3.16.** Partition de la sphère unité en méridiens et parallèles (projection orthographique). Le centre de chaque cellule définit une orientation prototype.

La TH 3D est définie comme une application $h$ de $\Xi \times \Theta \times \Psi$ dans $R$, construite comme suit. Tout d'abord, $h$ est initialisée à zéro. Ensuite, pour tout point $p$ de $E$ et pour tout $j$ de $\{0,1,\ldots,N_\theta - 1\}$ et $k$ de $\{0,1,\ldots,N_\varphi - 1\}$, on construit la famille des plans passant par $p$, d'orientation $(\theta_j, \varphi_k)$. Pour chaque plan, on calcule sa distance signée à l'origine du repère, et exprimée par :

$$s_{jk}^p = n_x(\theta_i,\varphi_k)x_p + n_y(\theta_i,\varphi_k)y_p + n_z(\theta_i,\varphi_k)z_p . \tag{3.54}$$

Si $s_{jk}^p$ prend une valeur positive, elle est quantifiée à la plus proche valeur $\hat{s}_{jk}^p$ de $\Xi$, et l'accumulateur $h(\hat{s}_{jk}^p, \theta_j, \varphi_k)$ est incrémenté de $w_{jk}^p$, donnant la contribution du point p sur le plan d'orientation

( $\theta_j$ , $\varphi_k$ ). En revanche, si $s_{jk}^p$ prend une valeur négative, l'accumulateur reste inchangé. Ce test sur le signe de $s_{jk}^p$ est une des solutions possibles pour éviter de générer des TH redondantes. En effet, si l'on considère l'ensemble symétrique $\tilde{\Xi} = \{ s_k = k\Delta_s \,|\, k \in \{-N_s+1, \ldots, -1, 0, 1, \ldots N_s - 1\} \}$, la TH associée, notée $\tilde{h}$ , vérifie :

$$\tilde{h}(s, \theta, \varphi) = \tilde{h}(-s, \theta + \pi, -\varphi), \tag{3.55}$$

ce qui exprime qu'à un même plan de l'espace, on peut associer deux vecteurs orientation ayant la même direction et des sens opposés. Les contributions accumulées sur chacun de ces deux plans sont bien évidemment égales, d'où la redondance introduite. Notre solution revient à sélectionner l'unique plan correspondant à une distance signée positive, par rapport à l'origine du repère.
Observons qu'il reviendrait au même, en terme de compacité des TH construites, que les prototypes d'orientation associés recouvrent uniquement une demi-sphère, empêchant ainsi la coexistence d'orientations de sens opposés dans l'ensemble des prototypes de quantification.

Dans le cas des maillages polygonaux, nous prenons pour *E* l'ensemble des centres de gravité de toutes les facettes du maillage. Quant au choix des contributions $w_{jk}^p$ de chaque facette, plusieurs variantes sont possibles, comme nous allons en discuter à présent.

### 3.3.1.2 Pondération par critère d'orientation

Une première possibilité consiste à prendre $w_{jk}^p$ égal à l'aire $A_p$ de la facette *p* du maillage, quels que soient *j* et *k*. Cela a l'avantage de prendre en compte l'aire des facettes et de pouvoir donc gérer des modèles irrégulièrement maillés. En revanche, l'information d'orientation des facettes est complètement négligée, ce qui conduit, conjointement aux effets de quantification des paramètres ($s$, $\theta$, $\varphi$), à une sur-accumulation d'aires "parasites".
Pour pallier cet inconvénient, une solution consiste à choisir des contributions dépendant non seulement des aires mais également des orientations des facettes. Nous proposons donc de définir $w_{jk}^p$ comme suit :

$$\text{(H1)} \qquad w_{jk}^p = A_p \delta(\theta^p - \theta_j, \varphi^p - \varphi_k),$$

où $\theta^p$ et $\varphi^p$ spécifient l'orientation de la facette *p* (après quantification dans $\Theta$ et $\Psi$). Cette formulation (H1) présente cependant l'inconvénient d'être trop fortement influencée par les orientations des facettes, puisqu'il existe des maillages présentant le même aspect mais pour des orientations de facettes très différentes. Pour limiter ce biais, nous considérons l'information d'orientation de manière moins fruste, selon la formulation (H2) :

$$\text{(H2)} \qquad w_{jk}^p = A_p \left| \left\langle n^p, n_{jk} \right\rangle \right|,$$

où $n^p$ désigne le vecteur normal à la facette *p*, $n_{jk} = (\cos(\varphi_k)\cos(\theta_j),\ \cos(\varphi_k)\sin(\theta_j),\ \sin(\varphi_k))^t$, $\langle \cdot, \cdot \rangle$

est le produit scalaire entre deux vecteurs de $\mathbb{R}^3$ et $T \in [0,1]$ est un seuil à fixer.

Dans l'hypothèse (H2), $w_{jk}^p$ est en effet égal à l'aire de la projection de la facette considérée sur le plan d'orientation ($\theta_j$, $\varphi_k$).
Nous proposons d'unifier les deux formulations précédentes au sein d'un même schéma hybride (H3) :

$$\text{(H3)} \qquad w_{jk}^p = \begin{cases} A_p \left| \left\langle n^p, n_{jk} \right\rangle \right|, & si \left| \left\langle n^p, n_{jk} \right\rangle \right| \geq T \\ 0, & \text{sinon} \end{cases},$$

les variantes (H1) et (H2), devenant des cas particuliers de (H3), lorsque $T \to 1$ et $T = 0$, respectivement.

Par la construction de la transformée de Hough, les maxima de la fonction *h* correspondent aux éléments planaires prédominants (au sens de la quantification considérée de l'espace des paramètres) constituant l'objet.
Toutefois, en pratique, cela n'est valable qu'après adaptation de la taille des facettes du maillage à la granularité de la transformée. Il est donc nécessaire d'effectuer un ré-échantillonnage adaptatif du maillage, comme décrit au paragraphe suivant.

### 3.3.1.3 Granularité et remaillage adaptatif

Le principe de construction de la TH présentée ci-dessus repose implicitement sur l'hypothèse que les tailles des facettes individuelles du maillage sont bien adaptées à la granularité de la transformée, définie par le pas de quantification $\Delta_s$ des distances des facettes à l'origine du repère considéré.

Pour illustrer concrètement ce propos, considérons un maillage constitué d'une unique facette, de centre de gravité $p = (x_p, y_p, z_p)$ et d'orientation $n = (n_x, n_y, n_z)$, et appliquons lui une TH dans la variante où $w_{jk}^p$ est égal à $A_p$, quels que soient *j* et *k*. Le lieu géométrique des points de l'espace de Hough recevant des contributions non nulles est le graphe d'une fonction sinusoïdale en $\theta$ et $\varphi$, donnée par l'équation suivante :

$$s(\theta, \varphi) = x_p \cos\theta \cos\varphi + y_p \cos\theta \cos\varphi + z_p \sin\varphi \ . \tag{3.56}$$

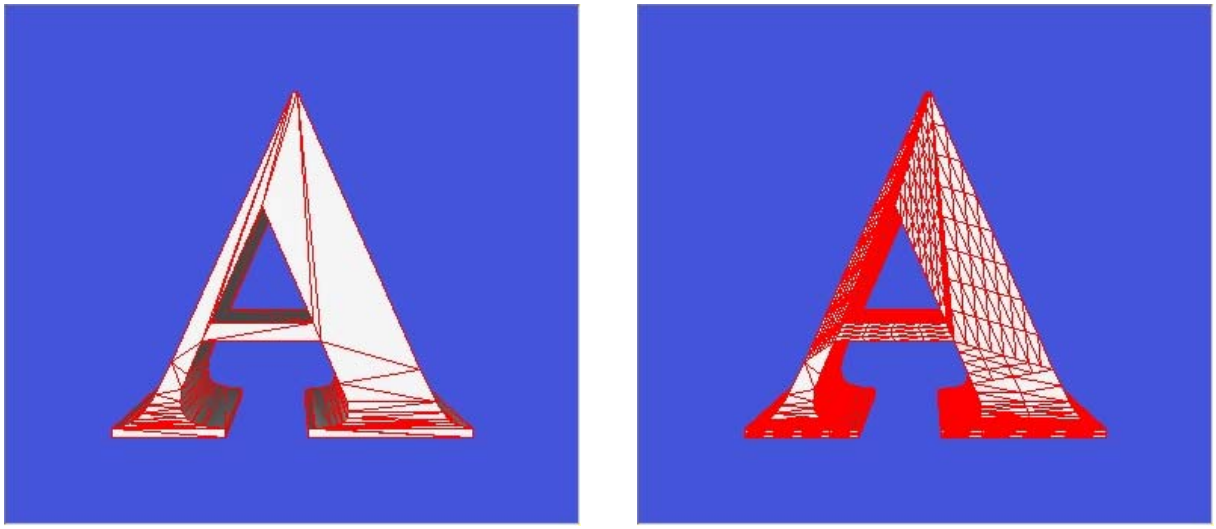
En outre, ces contributions sont identiques et égales à l'aire de la facette considérée, quels que soient $\theta$ et $\varphi$. En conséquence, il est impossible de remonter à l'information de localisation et d'orientation de la facette à partir de sa TH.

Considérons maintenant la même facette, mais subdivisée de manière suffisamment dense en *F* facettes distinctes, de centres de gravités notés $\left\{ x_p^i, y_p^i, z_p^i, \right\}_{i=1}^F$. En lui appliquant la TH, chaque facette *i* donnera une contribution non nulle et égale à son aire sur le graphe $\Gamma_i$ d'une fonction de la forme :

$$s^i(\theta, \varphi) = x_p^i \cos\theta \cos\varphi + y_p^i \cos\theta \cos\varphi + z_p^i \sin\varphi \ . \tag{3.57}$$

Les graphes $\Gamma_i$ s'intersectent en un unique point de l'espace de Hough, noté ($s_p$, $\theta_p$, $\varphi_p$), où $s_p$ est la distance de l'orgine du repère à la facette, et ($\theta_p$, $\varphi_p$) sont les angles sphériques associés à son orientation *n*. La TH est maximale en ce point et sa valeur égale à l'aire de la facette. Ce maximum permet donc de déterminer complètement le plan associé à la facette, par son orientation et sa distance à l'origine du repère.

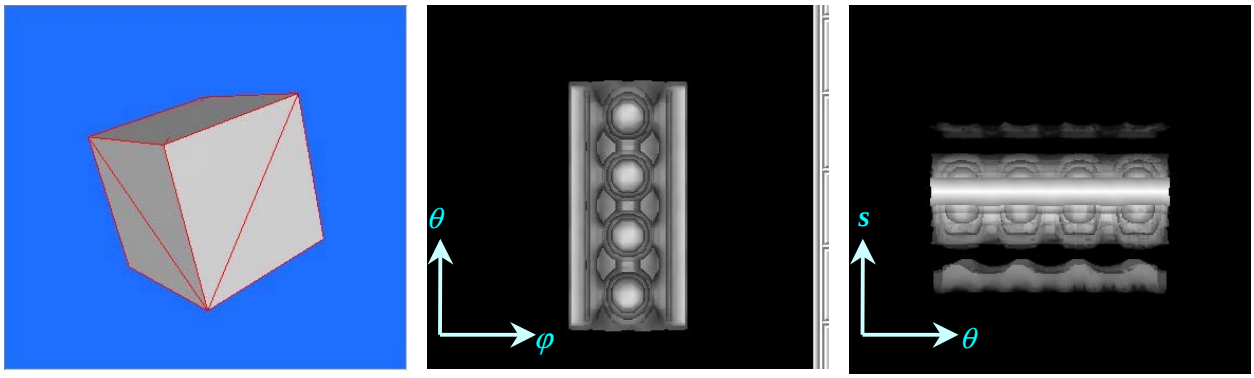
En d'autres termes, cela signifie que les arêtes des facettes du maillage doivent être du même ordre de grandeur que $\Delta_s$, qui s'exprime à sont tour en fonction de la taille $S_{max}$ du maillage et du nombre d'intervalles de quantification $N_s$.
En pratique, les maillages quelconques ne respectent en général pas cette hypothèse, pouvant être constitués de facettes de taille variable et même comparables à la taille $S_{max}$ du maillage (Figure 3.17.a).
Une façon simple de satisfaire à cette contrainte est d'appliquer une subdivision récursive des facettes n'y satisfaisant pas (Figure 3.17.b). Notons qu'alors l'information de connexité est irréversiblement dégradée, ce qui est sans conséquence puisque le descripteur que nous allons construire sera indépendant de celle-ci.



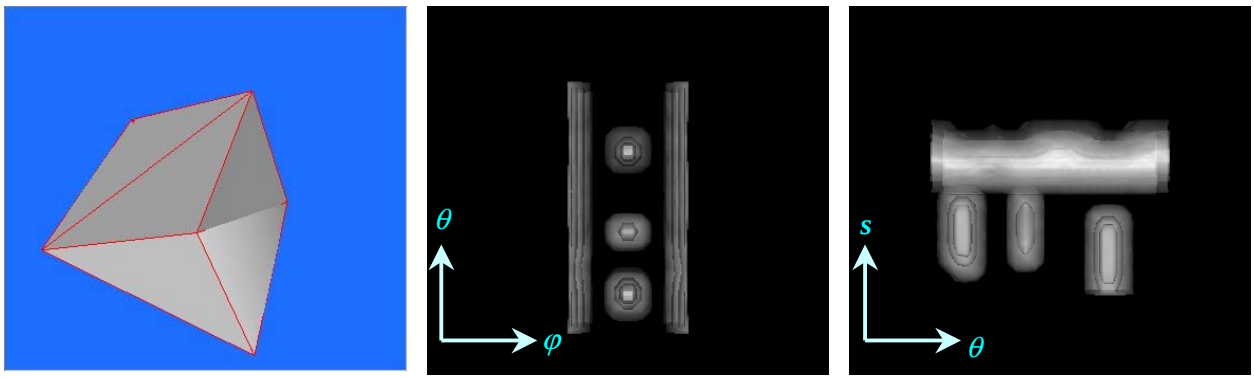
a. Modèle initial. b. Version remaillée.

**Figure 3.17.** Remaillage adaptatif des modèles polygonaux.

Une fois ces détails d'ordre pratique complètement spécifiés, illustrons tout d'abord le comportement de la TH sur quelques exemples simples, tels qu'un cube et un prisme triangulaire (Figure 3.18).




a. Cube triangulé.




b. Prisme triangulé.

**Figure 3.18.** Formes élémentaires et transformées de Hough 3D
(par rendu surfacique en transparence) visualisées dans les repères ($\varphi$, $\theta$) et ($\theta$, $s$).

Dans le cas du cube, six maxima, un par facette sont bien visibles. Leur localisation est parfaitement équidistante selon les axes $\theta$ et $\varphi$ et leurs distances à l'origine sont identiques. Cela n'est plus le cas pour le prisme triangulaire, où les 5 maxima correspondant aux 5 facettes sont différemment distribués, aussi bien selon l'axe des s, que selon celui des $\theta$.

La TH3D offre donc une signature de forme discriminante. Etudions à présent comment il est possible d'en dériver un DF optimisé en compacité.

## 3.3.2 Le descripteur de Hough 3D optimisé

Tous les développements précédents sont effectués dans le cadre d'un repère cartésien connu. Afin d'obtenir un descripteur invariant par rapport aux translations et homothéties, un repère lié à l'objet doit tout d'abord être déterminé. Cela revient à résoudre le problème d'alignement.

### 3.3.2.1 Alignement spatial : les configurations génératrices

La construction du repère local s'appuie sur une analyse en composantes principales (ACP). Le centre de gravité (surfacique) du maillage est choisi comme origine du repère. La matrice de covariance $\sigma$ des coordonnées des centres de gravité des facettes est ensuite calculée. Cette matrice semi-définie positive

est donc diagonalisable par une transformation orthogonale appropriée $R$. En outre, ses valeurs propres sont réelles et positives et donnent les longueurs des axes de l'ellipsoïde d'inertie associé au modèle. Ces trois valeurs propres, $\lambda_1$, $\lambda_2$ et $\lambda_3$, indiquent également la dimension de l'objet.
Posant

$$S_{\max} = 1.5\sqrt{\lambda_1 + \lambda_2 + \lambda_3}\ , \tag{3.58}$$

nous définissons une échelle intrinsèque à l'objet pour laquelle seules les facettes y satisfaisant interviennent dans le calcul du descripteur (les facettes périphériques ne vérifiant pas ce critère ne sont tout simplement pas prises en compte).

Pour aligner spatialement les objets, la plupart des méthodes de la littérature consistent à étiqueter les trois axes principaux par ordre croissant ou décroissant des valeurs propres. Ce choix se révèle bien souvent insuffisant. Tout d'abord, l'ACP ne permet pas de déterminer le sens des axes : des analyses complémentaires sont alors nécessaires. De plus, dans certaines situations, la représentation des modèles dans ces repères conduit à de faux alignements.
Pour illustrer ce propos, considérons les modèles d'avion présentés Figure 3.19. L'axe d'inertie de la plus grande valeur propre pour le modèle de la Figure 3.19.a suit la direction du fuselage de l'avion, tandis que dans le cas du modèle de la Figure 3.19.b, il est orienté selon la direction des ailes. Il en résulte un alignement erroné.

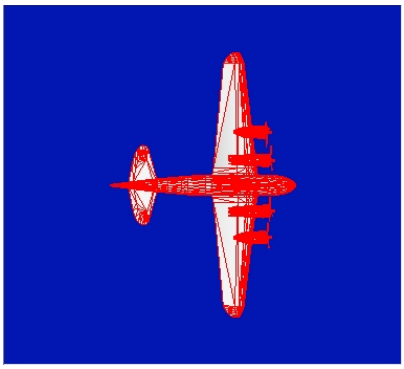
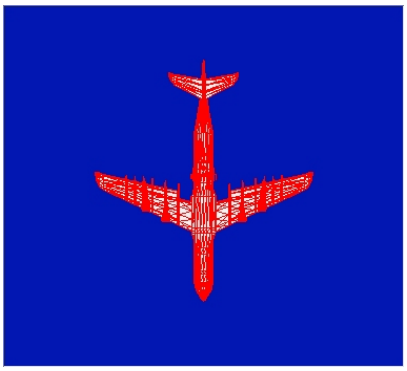
a.   b.

**Figure 3.19.** Exemple d'alignement erroné.
L'axe principal de plus grande valeur propre correspond à la verticale.

En raison de ces difficultés, nous conservons la totalité des représentations possibles correspondant à tous les repères que l'on peut construire avec des axes parallèles aux axes principaux de l'objet. En considérant l'étiquetage ($x$, $y$, $z$) des trois axes principaux avec deux sens possibles pour chaque axe, il y a $48 = 6 \times 4 \times 2$ manières différentes de définir un repère. Mais retenir les 48 représentations correspondant à ces 48 repères conduirait à une complexité trop élevée du descripteur. Nous proposons de la réduire drastiquement en exploitant les relations existant entre ces 48 représentations.

Appelons *configurations génératrices*, notées $h_1$, $h_2$ et $h_3$, les trois transformées de Hough 3D de l'objet associées aux trois repères canoniques, définis comme suit. Chacun des trois axes d'inertie de l'objet devient successivement l'axe des $z$ du repère, avec un sens arbitraire.

Montrons qu'il suffit de ces trois configurations génératrices pour disposer d'une représentation complète au sens où les autres représentations associées aux autres choix de repère s'en déduisent mathématiquement.

Considérons les 16 repères possibles pour un choix fixé de la direction de l'axe *z*. Pour chaque sens de l'axe *z*, on obtient 8 repères différents, déduits par rotations successives de 90° et réflexions miroir (Figure 3.20) du repère initial.

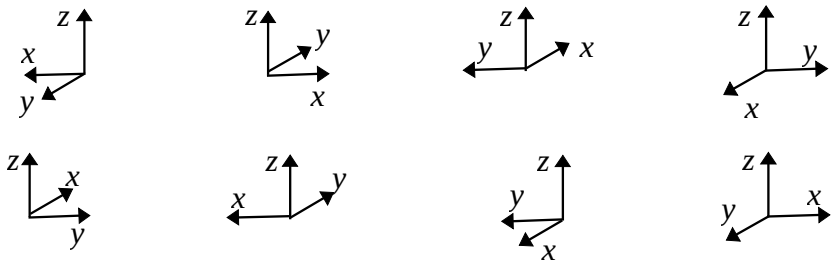



**Figure 3.20.** Les 8 repères différents obtenus pour un axe *z* fixé.

Ces transformations géométriques se traduisent sur la TH en termes de translations circulaires et de réflexions miroir des coefficients. Notons par $h^0$ la TH du modèle initial et par $h$ celle du modèle transformé. Une rotation à 90° autour de l'axe de *z* s'écrit :

$$h(s_i, \theta_j, \varphi_k) = h^0(s_i, \theta_{<j+\frac{N_\theta}{4}>_{N_\theta}}, \varphi_k), \tag{3.59}$$

où $<\cdot>_c$ désigne *modulo c* et $N_\theta$ est supposé être multiple de 4.

L'inversion du sens de l'axe *x* est donnée par :

$$h(s_j, \theta_j, \varphi_k) = h^0(s_j, \theta_{<N_\theta - j>_{N_\theta}}, \varphi_k). \tag{3.60}$$

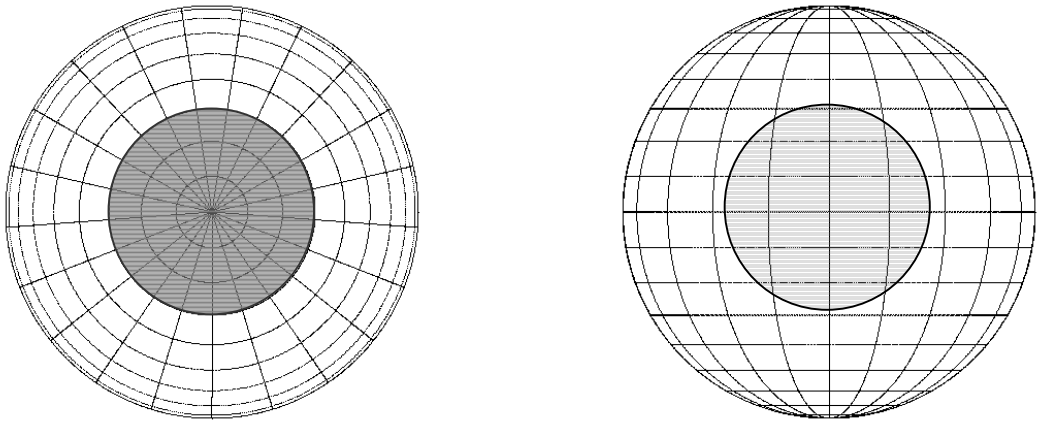
De façon analogue, l'inversion du sens de l'axe de *z* s'exprime par :

$$h(s_j, \theta_j, \varphi_k) = h^0(s_j, \theta_j, \varphi_{<N_\varphi - k>_{N_\varphi}}). \tag{3.61}$$

Le caractère générateur de $\{h_1, h_2, h_3\}$ étant établi, poursuivons en étudiant son caractère minimal, au sens où l'on peut trouver ou non une bijection permettant de déduire une configuration génératrice à partir des deux autres.

### 3.3.2.2 Partition polyédrique régulière de la sphère : le DH3D optimisé

L'échantillonnage uniforme des coordonnées sphériques $\theta$ et $\varphi$ conduit à des partitions en "parallèles" et "méridiens" de la sphère unité, différentes dans les trois repères canoniques considérés, comme illustré Figure 3.21.



a. b.

**Figure 3.21.** Non-équivalence des partitions de la sphère unité par changement de repère canonique.

En effet, les cellules de cette partition illustrées Figure 3.21.a ne se recalent pas sur les cellules correspondant à un autre repère canonique. En conséquence, $\{h_1, h_2, h_3\}$ est un ensemble générateur minimal, en raison de cette discrétisation liée au repère. Dans ce cas, le DH3D associé à la paramétrisation sphérique doit prendre en compte les trois configurations génératrices $h_1$, $h_2$, $h_3$, qui le définissent complètement.

Néanmoins, en abandonnant la partition par paramétrage sphérique, il est possible de définir des partitions dans lesquelles les configurations génératrices deviennent équivalentes, via une bijection. En considérant par exemple les partitions de la sphère unité obtenues par projection des sommets d'un polyèdre régulier sur la sphère, l'invariance aux trois choix possibles de repère canonique est assurée. Notons que des partitions aussi fines que l'on veut sont obtenues en appliquant par exemple un mécanisme de subdivision des facettes du polyèdre, comme illustré Figure 3.22 pour le cas d'un octaèdre.

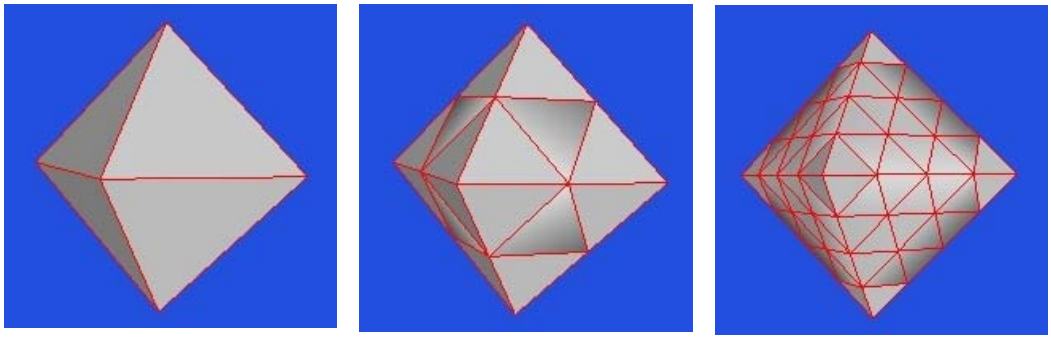


**Figure 3.22.** Subdivisions d'un octaèdre.

Ainsi, nous pouvons construire un DH3D optimal (DH3DO) en terme de compacité, spécifié par une unique transformée de Hough, associée à l'un des 48 repères possibles.

#### 3.3.2.3 Mesures de similarité

Pour la mise en œuvre du DH3DO, les mesures de similarité considérées sont les distances $L_1$ ou $L_2$ dans l'espace des coefficients de Hough et le schéma de principe est le suivant :

- Considérer le DH3DO du modèle requête, noté $m_\rho$.
- Générer, à partir du DH3DO de chaque modèle de la base $m_b$, les 48 TH 3D correspondant à l'ensemble des repères considérés, notées $\left\{h_b^k\right\}_{k=1}^{48}$, puis calculer leurs distances au DH3DO du modèle requête, exprimées par :

$$\delta_k(m_\rho, m_b) = \delta(h_\rho, h_b^k), \tag{3.62}$$

  où $h_\rho$ est la TH du modèle requête telle que stockée dans le DH3DO et $\delta(\cdot,\cdot)$ la distance $L_1$ ou $L_2$.

- Définir comme mesure de similarité la distance minimale parmi les 48 distances calculées, sous la forme :

$$\Delta(m_\rho, m_b) = \min_{k\in\{1,2,\ldots,48\}} \left\{\delta(h_\rho, h_b^k)\right\}. \tag{3.63}$$

Il est intéressant de remarquer que la mesure de similarité ainsi définie est naturellement symétrique. Cette propriété est intrinsèquement liée à la partition invariante au changement de repère canonique de la sphère unité, qui permet d'exprimer les relations existant entre les configurations génératrices par des simples permutations des coefficients. Remarquons que cette propriété n'est pas du tout valable dans le cas des partitions par échantillonnage uniforme des coordonnées sphériques, où, en raison de la non-équivalence des configurations génératrices, l'approche décrite ci-dessus conduirait à des distances non-symétriques. Ce problème peut être resitué dans un contexte plus général, car il est intrinsèquement associé à tout descripteur à base de représentations multiples et non équivalentes, comme par exemple dans le cas du *Multiview DS* MPEG-7 (*cf.* Chapitre 1). Symétriser alors les mesures de similarité est effectivement possible (en considérant par exemple des distances de type Haussdorff), la contrepartie étant l'augmentation significative du coût de calcul, qui devient quadratique par rapport au nombre de représentations.

Notons enfin qu'il est possible de réduire encore davantage le coût du DH3DO. Par exemple, les valeurs propres donnant les longueurs de l'ellipsoïde d'inertie de l'objet peuvent être exploitées pour éviter de calculer les distances entre des TH correspondant à des non-alignements flagrants. Ces aspects ne seront cependant pas abordés ici.

## 3.4 Evaluation expérimentale des descripteurs de forme

Les résultats expérimentaux ont été conduits sur la base de test MPEG-7 composée de 1300 modèles au format VRML 2.0. Présentons donc cette base et la catégorisation associée.

### 3.4.1 Les corpus d'étude

Ces modèles proviennent de trois corpus différents :

- 293 modèles utilisés par le sous-groupe MPEG-SNHC (*Synthetic and Natural Hybrid Coding)*, dans le cadre des développments des technologies de compression MPEG-4 de maillages 3D [MPEG-4],
- 50 modèles correspondant aux lettres de "A" à "E", avec pour chaque lettre 10 maillages différents, spécialement créés pour tester la robustesse des descripteurs par rapport aux problèmes de représentations topologiques multiples [Murao-Iso99.10],
- une sélection de 947 modèles de la base *3D Cafe* [3DCafe], disponible sur *Internet* et retenue à notre instigation comme base de test MPEG-7.

Afin d'établir une "vérité terrain", indispensable pour l'évaluation quantitatives des performances des descripteurs, un sous-ensemble de 362 modèles a été catégorisé (Tableau 3.1), représentant l'extension d'une première catégorisation sur 228 modèles proposée initialement par les auteurs et acceptée dans le cadre des expérimentations techniques de MPEG-7 sur le SF3D.

| Catégorie | Q | Catégorie | Q |
|---|---|---|---|
| Avions | 52 | Tournevis | 9 |
| Humanoïdes | 24 | Pièces cylindriques | 28 |
| Voitures | 37 | Arbres sans feuilles | 3 |
| Chars d'assaut | 6 | Arbres avec feuilles | 23 |
| Camions | 9 | Pièces sphériques | 27 |
| Formule 1 | 11 | Doigts | 30 |
| Motos avec attache | 3 | Lettre "A" | 10 |
| Motos | 10 | Lettre "B" | 10 |
| Hélicoptères | 9 | Lettre "C" | 10 |
| Pistolets | 12 | Lettre "D" | 10 |
| Fusils | 10 | Lettre "E" | 10 |
| Pièces d'échec | 10 | | |

**Tableau 3.1.** Les catégories retenues. Q désigne le nombre de modèles par catégorie.

Notons que cette catégorisation n'est pas purement sémantique, des modèles sémantiquement différents mais géométriquement similaires ayant été regroupés dans la même catégorie. Malgré cela, une variabilité importante en terme de perception intuitive de forme persiste au sein de la plupart de ces 23 catégories : des avions à une ou plusieurs ailes, simples ou en delta, des humanoïdes avec des positions de bras et jambes bien différentes, des voitures anciennes ou modernes, décapotables ou non sont respectivement regroupés dans la classe générique "Avions", "Humanoïdes", "Voitures" (Figure 3.23-Figure 3.25).
Cela rend la base catégorisée suffisamment complexe pour tester effectivement les performances d'un DF 3D.

**Figure 3.23.** Quelques modèles de la catégorie "Avions".

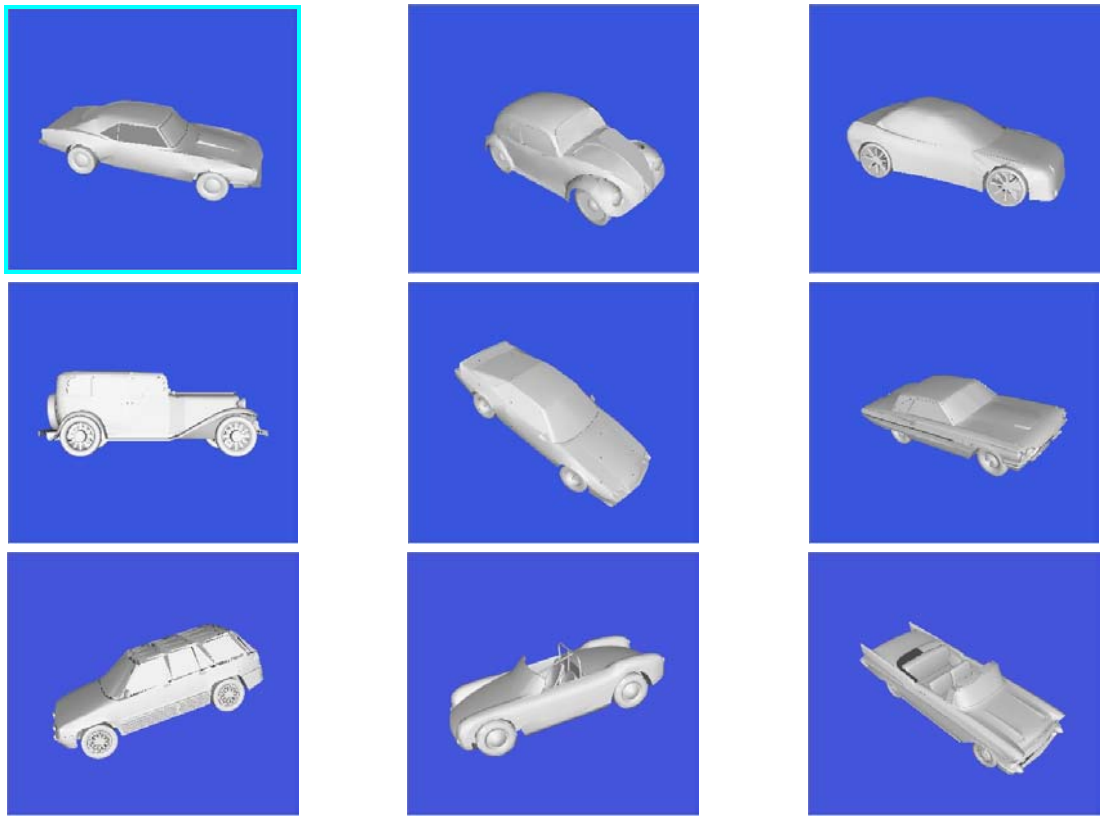


**Figure 3.24.** Quelques modèles de la catégorie "Voitures".

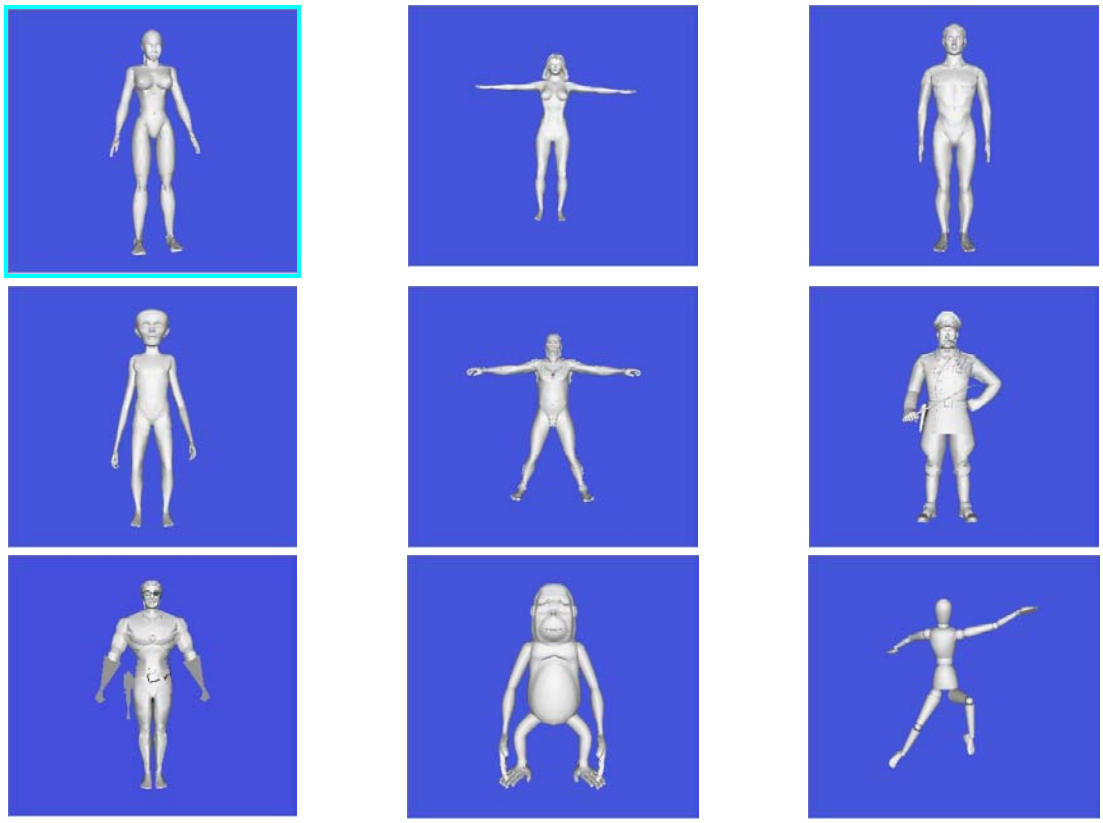


**Figure 3.25.** Quelques modèles de la catégorie "Humanoïdes".

### 3.4.2 Résultats et analyse comparée

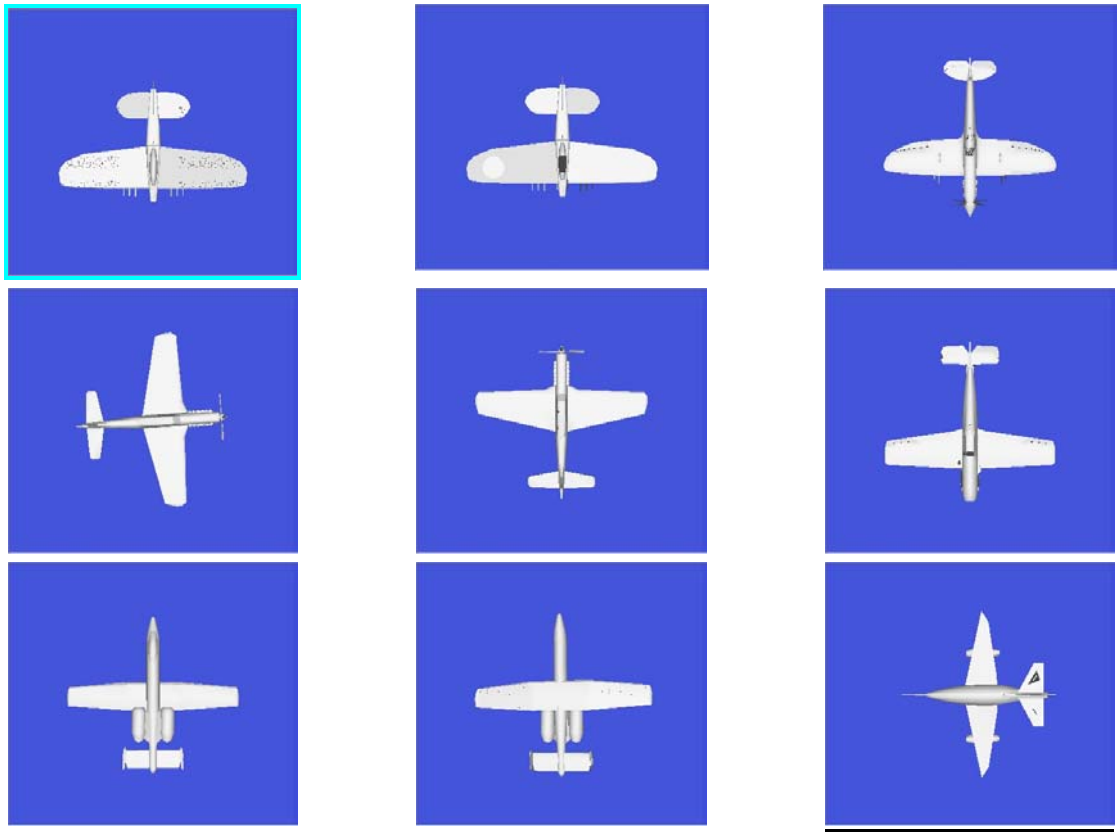
Les résultats présentés par la suite sont obtenus avec $N_{bins}$ = 100 et une représentation en virgule flottante pour les valeurs d'histogramme, dans le cas du SF3D, et $N_s$ = 20, une partition sur deux niveaux de subdivision d'un octaèdre et une pondération (H3) avec $T$ = 0.7, dans le cas du DH3DO.

Pour illustrer les performances des requêtes par similarité des deux descripteurs proposés, présentons un premier exemple de requête emprunté à la catégorie "Avions" (Figure 3.26). Les modèles retrouvés sont triés par ordre décroissant de similarité par rapport au modèle requête, de gauche à droite et de haut en bas.
Notons l'apparition de résultats erronés retrouvés aux 3$^e$, 4$^e$ et 9$^e$ positions, lorsque le SF3D est utilisé. Cela ne se produit pas avec le DH3DO, pour lequel des avions similaires à celui de la requête sont retrouvés aux 9 premières positions.

Un deuxième exemple emprunté à catégorie "Humanoïdes" (Figure 3.27) montre les limitations du DH3DO. En effet, aux neuf premières places, deux résultats erronés sont retrouvés par le DH3D, alors que le SF3D réussit à ne retrouver que des humanoïdes en dépit des différences fortes dans les positions des bras et des jambes.

Enfin, un dernier exemple, correspondant à une requête de la catégorie "Voitures", est présenté Figure 3.28. Une fois encore, le DH3D montre ici un meilleur comportement que celui du SF3D.

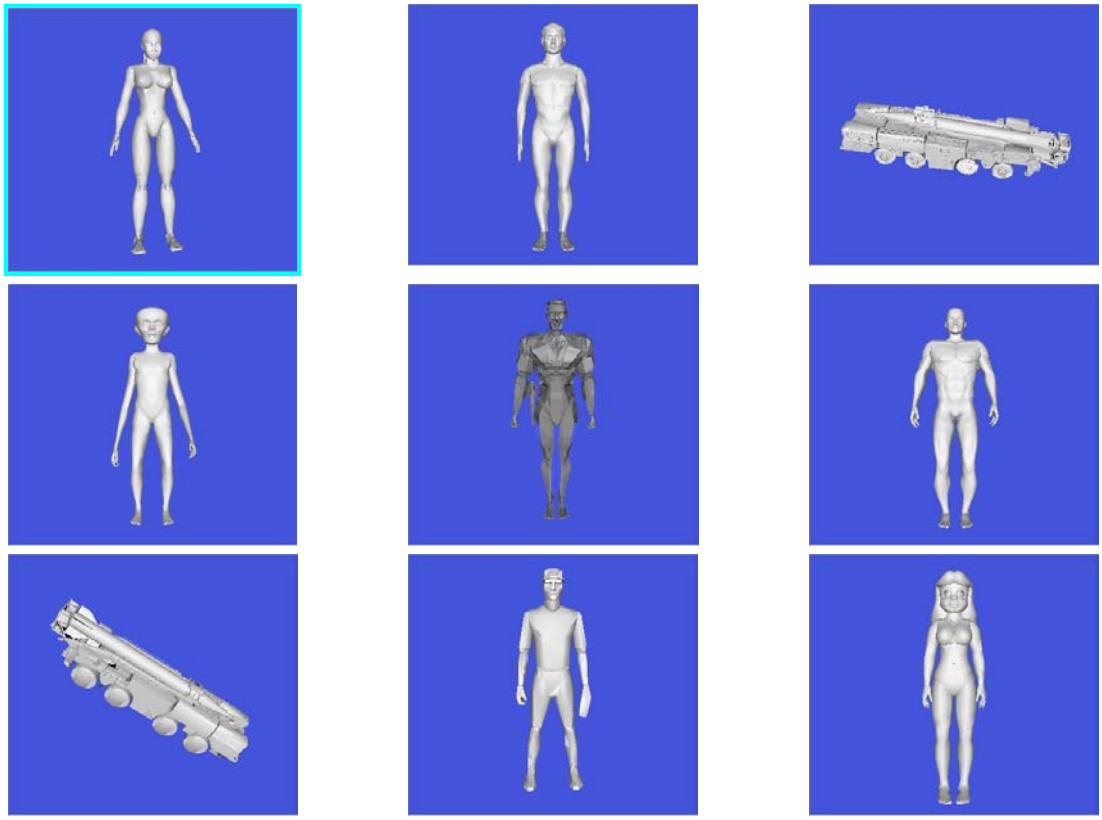
a. SF3D



b. DH3DO.

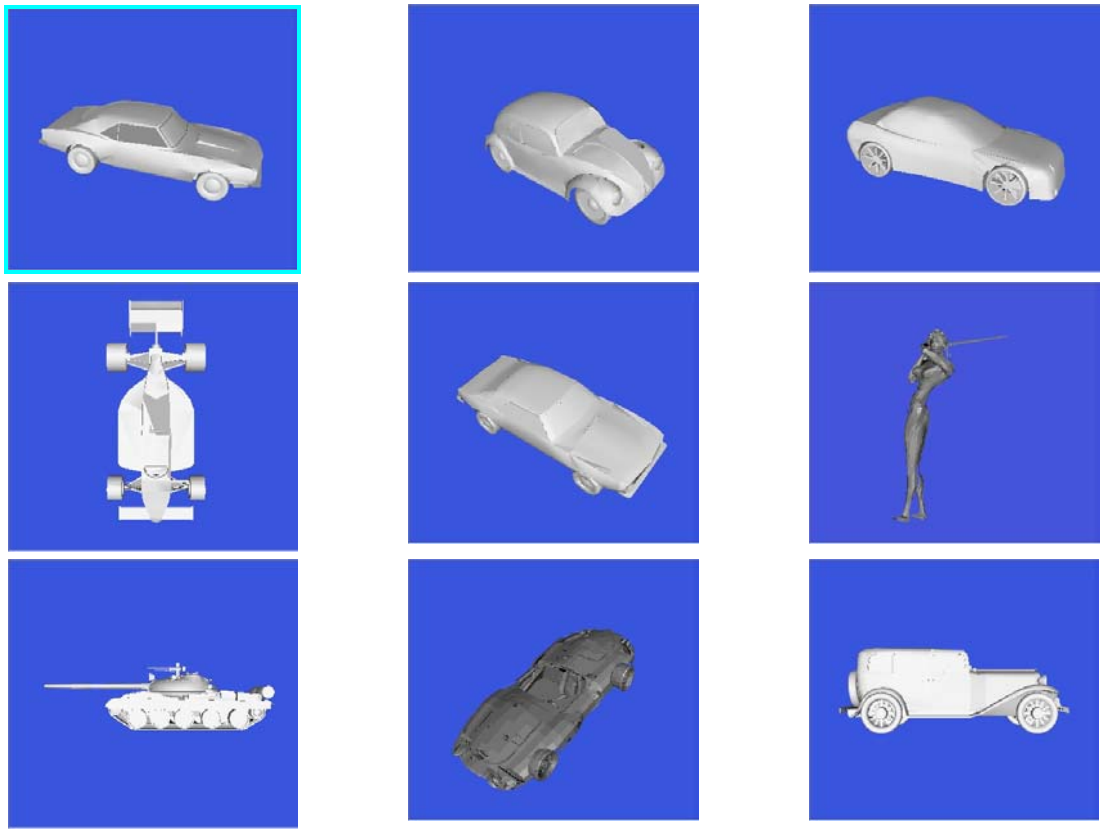
**Figure 3.26.** Modèles retrouvés pour une requête de la catégorie "Avions"

a. SF3D

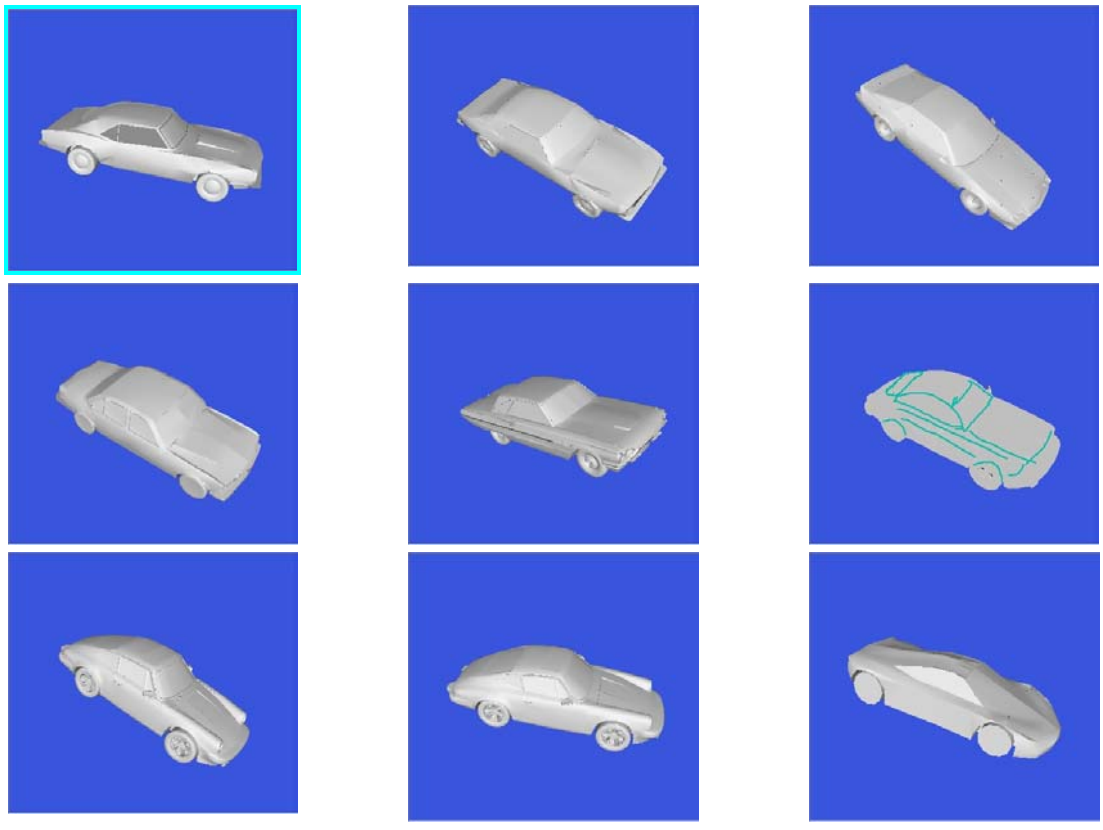


b. DH3DO.

**Figure 3.27.** Modèles retrouvés pour une requête de la catégorie "Humanoïdes".



a. SF3D.



b. DH3DO.

**Figure 3.28.** Modèles retrouvés pour une requête sur la catégorie "Voitures".

Pour évaluer quantitativement et objectivement les performances des DF proposés, nous avons calculé le score *Bull-Eye* (SBE) sur toutes les catégories définies. Pour chaque maillage catégorisé, le SBE est défini comme le pourcentage de résultats corrects (*i.e.* maillage appartenant à la même catégorie que celle de la requête) parmi les 2Q premiers résultats retrouvés, où Q est le nombre d'éléments de la catégorie correspondant à la requête. Pour chaque modèle maillé de chaque catégorie, le SBE est calculé, puis le SBE moyen par catégorie est estimé (Tableau 3.2). Mentionnons que le critère SBE a été la base d'évaluation de tous les descripteurs de forme considérés dans MPEG-7.

En terme de scores moyens globaux, sur l'ensemble des catégories (Tableau 3.3), l'excellent score du DH3DO (81%) offre un gain significatif de 28% par rapport à celui du SF3D (53%), démontrant sa nette supériorité également pour chaque catégorie. Cette performance est étroitement liée d'une part à une représentation globale plus complète de l'information de forme et d'autre part aux propriétés d'invariance géométrique et topologique naturelles ou conférées au DH3DO. En outre, le mécanisme de pondération (H3) contribue à l'excellent score obtenu, comme l'établit l'expérimentation conduite en estimant le DH3DO sans prendre en compte les orientations des facettes ($\forall\ j,k,\ w_{jk}^{p} = A_p$ ), conduisant à un SBE qui tombe à 76%.

| Catégorie | Q | SF3D | H3D | Catégorie | Q | SF3D | H3D |
|---|---|---|---|---|---|---|---|
| Avions | 52 | 66 | 94 | Tournevis | 9 | 49 | 90 |
| Humanoïdes | 24 | 56 | 63 | Pièces cylindriques | 28 | 34 | 87 |
| Voitures | 37 | 44 | 85 | Arbres sans feuilles | 3 | 55 | 100 |
| Chars d'assaut | 6 | 16 | 55 | Arbres avec feuilles | 23 | 43 | 66 |
| Camions | 9 | 26 | 55 | Pièces sphériques | 27 | 42 | 97 |
| Formule 1 | 11 | 17 | 72 | Doigts | 30 | 100 | 100 |
| Motos avec attache | 3 | 11 | 100 | Lettre "A" | 10 | 100 | 100 |
| Motos | 10 | 43 | 49 | Lettre "B" | 10 | 100 | 100 |
| Hélicoptères | 9 | 35 | 64 | Lettre "C" | 10 | 100 | 100 |
| Pistolets | 12 | 21 | 52 | Lettre "D" | 10 | 100 | 100 |
| Fusils | 10 | 29 | 40 | Lettre "E" | 10 | 100 | 100 |
| Pièces d'échec | 10 | 31 | 88 | | | | |

**Tableau 3.2.** SBE moyens (%) par catégorie.

| | SF3D | H3D |
|---|---|---|
| ***SBE global (%) :*** | 53 | 81 |

**Tableau 3.3.** SBE globaux sur l'ensemble des catégories.

Mentionnons également la supériorité du DH3DO par rapport au SF3D en terme de scalabilité (i.e. persistance de bonnes performances avec l'augmentation du volume de la base de données). Ainsi, lorsque l'ensemble des modèles catégorisés est replongé dans la totalité de la base des 1300 modèles 3D, le DH3DO fournit un SBE moyen global de 75%, tandis que le SF3D chute à 43%.

Les performances supérieures du DH3DO par rapport au SF3D en termes de performances de requêtes par similarité sont en revanche pénalisées par une complexité de calcul largement plus importante (48 distances entre vecteurs dans $\mathbb{R}^{1028}$ à calculer pour le DH3DO avec deux niveaux de subdivision de la sphère unité, par rapport à une seule distance entre deux vecteurs de $\mathbb{R}^{100}$ pour le SF3D, avec une représentation habituelle sur 100 intervalles de quantification).
Afin de réduire encore plus la complexité du descripteur en termes de stockage et temps de réponse aux requêtes, nous avons appliqué au DH3DO une transformée de Fourier 2D discrète. La partition de la sphère unité par subdivisions du tétraèdre ne permettant pas de définir de manière directe une vraie transformée de Fourier 3D, l'intégration a été réalisée indépendamment selon les axes $\theta$ et $s$. Les valeurs absolues des coefficients de la transformée sont intrinsèquement invariantes aux rotations d'axe $z$ (qui se traduisent par des translations circulaires en $\theta$), ce qui permet de réduire le nombre de configurations à générer et à comparer, de quarante huit à douze. En utilisant une distance $L_1$ entre les coefficients de Fourier, de bons résultats sont encore obtenus, mais malgré tout avec une chute du SBE moyen sur l'ensemble des catégories à 69%.

## 3.5 Conclusion

Dans ce chapitre, nous avons présenté, analysé et comparé deux approches différentes pour représenter des modèles polygonaux 3D, visant des applications de requête par similarité de forme.
Partant des propriétés d'invariance géométrique et topologique que doit satisfaire naturellement un DF d'objet 3D maillé, nous avons proposé tout d'abord le SF3D. Celui-ci exploite uniquement la structure géométrique locale d'une surface 3D, fournit une représentation très compacte mais offre des performances moyennes en terme de requête par similarité en raison d'une part de la perte de toute information de localisation spatiale et d'autre part de sa grande sensibilité aux descriptions topologiques des maillages.
Ensuite, en considérant la transformée de Hough 3D d'un maillage, nous avons construit le DH3DO, intrinséquement invariant aux changements de connexité, rendu indépendant des transformations géométriques et optimisé en terme de compacité de représentation, via une partition invariante aux changements de repère canonique de la sphère unité. Ses performances, excellentes en terme de SBE (81%) sur une base catégorisée de 362 modèles, se conservent sur l'ensemble des 1300 modèles de la base MPEG-7, confirmant les remarquables propriétés de scalabilité du DH3DO.

# Chapitre 4

# Plate-forme AMIS d'indexation vidéo compatible MPEG-7 et applications


## Résumé

*Ce dernier chapitre présente le système d'indexation multimédia AMIS (Advanced Multimedia Indexing System), qui offre, dans le cadre du standard MPEG-7, une plate-forme de visualisation/navigation/annotation des contenus audio-visuels naturels et synthétiques. Ce système, fondé sur une architecture modulaire extensible, intègre les divers types de média (images fixes, séquences d'images, documents vidéos, maillages 3D…) sous leurs différents formats et offre un riche ensemble d'outils de segmentation temporelle et spatio-temporelle et d'extraction des descripteurs MPEG-7 pour des applications de recherche par le contenu. En particulier, le principe d'interopérabilité devient effectif par la mise en œuvre des schémas de description MPEG-7 qui permettent de s'affranchir des problèmes de cohérence, d'héritage, de hiérarchie, de synchronisation et d'intégration de signatures hétérogènes.*

*L'ensemble logiciel ainsi réalisé démontre pour la première fois en grandeur réelle, pour des applications d'indexation multimédia, comme l'archivage vidéo, la vidéo cliquable et la reconnaissance des gestes en langue des signes française, le caractère effectivement opérationnel des schémas de description audio-visuels génériques, normalisés MPEG-7.*



### Mots Clef

*Système d'indexation, standard MPEG-7, moteur de recherche, navigation, visualisation, annotation, résumés, schémas de description, archivage vidéo, vidéo cliquable, langue des signes, reconnaissance de gestes.*

## 4.1 Synthèse bibliographique

Comme nous l'avons vu au Chapitre 1, MPEG-7 permet d'intégrer de nombreux médias différents (image fixe, vidéo, audio, texte, objet 3D), au sein d'un large éventail d'applications comme celles à base de requêtes par similarité dans de grandes bases vidéos jusqu'à des applications de *broadcasting*, de vidéo à la demande ou encore celles spécifiques aux environnements mobiles.
Toutefois, l'exploitation efficace des technologies standardisées requiert de disposer de systèmes appropriés permettant visualisation, navigation, annotation et requêtes sur les contenus multimédias.

De tels systèmes doivent donc être capables de :

- supporter un large éventail de descripteurs couvrant efficacement des attributs visuels de couleur, texture, forme et mouvement et permettant d'effectuer des requêtes par le contenu,
- gérer, représenter et stocker de larges bases de vidéos et de métadonnées associées,
- offrir des interfaces graphiques permettant la spécification et la personnalisation des requêtes (selon le type d'objet ou de descripteur, ...)
- intégrer de multiples types de média, comme les vidéos codées en différents formats, les images fixes, les données 3D,
- disposer d'outils de visualisation, navigation, création de tables des matières…

Comment ses différentes contraintes sont-elles prises en compte dans les systèmes d'indexation déjà existants, qu'ils soient commerciaux ou de laboratoires ?

Référence classique des systèmes d'indexation commerciaux, QBIC (*Query by Image Content*) [Niblack93, Flickner95] combine des descripteurs visuels de couleur, forme, texture et mouvement pour la requête par le contenu des images fixes et des vidéos. Les représentations aussi bien "pleine image" que par objet de forme arbitraire sont supportées. Les descripteurs QBIC de forme incluent des paramètres élémentaires de circularité et excentricité (*cf.* Paragraphe 1.3.5.3) [Jain89] et les invariants algébriques proposés dans [Taubin91]. Les informations de texture, sont prises en compte par un descripteur simple, à base des caractéristiques de Tamura [Tamura78], qui expriment des mesures de régularité, direction et contraste. La couleur est représentée par des moments jusqu'au deuxième ordre et des histogrammes de couleur, les mesures de similarité associées étant celles décrites Paragraphe 1.3.5.1. Pour caractériser une vidéo, le système inclut des outils de segmentation temporelle en plans et en scènes, de sélection de trames clé, de détection d'événements, de modélisation paramétrique de mouvement et de segmentation par le mouvement [Ayer95]. Le système offre également une interface graphique pour la segmentation semi-automatique à base de contours actifs. Les descripteurs sont organisés dans des structures d'indexation multidimensionnelles de type R*-tree [Beckman90] supportant des requêtes en SQL [SQL]. La visualisation est fondée sur des icônes et tables des matières (*storyboards*).
La composante audio, absente dans QBIC, est prise en compte dans le cadre d'un second système d'indexation développé par IBM, appelé CueVideo [Mahmoo00], qui inclut des techniques de détection et

de reconnaissance de texte et de parole conjointement à un descripteur de couleur de type distribution spatiale (*color layout*). Les mêmes techniques de segmentation en plans et en scènes que dans QBIC sont utilisées ici.

Le système VideoLogger de Virage [Hamampur97] est fondé sur une architecture modulaire incluant segmentation temporelle en plans et en scènes aussi que sélection de trames clé pour la construction de tables de matières. VideoLogger exploite également des techniques de détection et reconnaissance de visage. Quant à l'information sonore, elle est prise en compte par des procédures de reconnaissance de la parole, visant à convertir les paroles détectées en du texte exploitable pour des requêtes textuelles.
Les principaux avantages du système résident dans :

- sa conception modulaire, qui permet de réaliser des logiciels dédiés à des applications spécifiques par sélection et réorganisation des modules spécifiques,
- sa capacité à supporter différents formats vidéos "streamés", indispensables dans le monde de l'Internet,
- la possibilité d'utiliser le système conjointement avec des logiciels dédiés de bases de données comme Oracle [Oracle] et Informix [Informix].

Soulignons que la technologie Virage a été considérée lors du dévéloppement du moteur de recherche AltaVista [AltaVista] dans le cadre de l'outil PhotoFinder.

Une approche similaire est proposée avec le système Excalibur [Feder96], les outils d'analyse vidéo propriétaires incluant des représentations par la couleur, la texture et la forme. Cette technologie a été considérée lors de l'élaboration du moteur de recherche d'images de Yahoo [Yahoo].

Les systèmes expérimentaux sont issus de recherches et développements dans les laboratoires académiques parfois en partenariat avec un industriel.

Dans ce contexte, le système ImageMiner [Alshuth98], développé à l'Université de Bremen (Allemagne) en coopération avec IBM, inclut des descripteurs de couleur, texture et forme. La distribution spatiale de la couleur est décrite par les couleurs dominantes estimées dans des blocs carrés de l'image. Les attributs texturaux sont pris en compte à l'aide de mesures de régularité, granularité et direction fournies par un réseau neuronal ayant pour entrées les traditionnelles informations de cooccurrences définies dans [Haralick73]. L'information de forme est spécifiée en termes de contours polygonaux, construits automatiquement en appliquant un algorithme d'extraction de contours [Canny86, Korn88] et de connexion des points de contours [Zhang94].
Pour structurer la vidéo, les auteurs proposent tout d'abord un découpage en plans dans le domaine compressé exploitant les coefficients DC des blocs MPEG, puis un regroupement des plans en scènes, selon un critère de similarité visuelle et sous une contrainte temporelle [Yeung97]. Cette segmentation est finalement représentée sous forme d'un graphe. Le contenu de chaque plans individuel est résumé par une image mosaïque, obtenue par modélisation paramétrique à base d'un modèle perspectif planaire.

Le système supporte aussi bien des requêtes de bas niveau à partir des différents descripteurs individuels considérés, que des requêtes de haut niveau, par exemple de nature sémantique, en exploitant un ensemble de concepts préalablement définis, ou de nature combinatoire en associant plusieurs descripteurs selon un formalisme à base de grammaire de graphes [Fröhlich95].

Le système MARS [Huang96], développé à l'Université Urbana Champaign (Illinois), offre des solutions intéressantes pour gérer et combiner différents descripteurs dans un cadre unifié. Deux descripteurs sont dédiés aux représentations de couleur, l'un à base d'histogramme global (dans l'espace HSV) et l'autre à base de distribution spatiale des couleurs dominantes estimées par un algorithme de k-moyennes. L'information de texture est représentée par les caractéristiques de Tamura et celle de forme par descripteurs de Fourier [Zahn72, Persoon77]. Les représentations par objets de forme arbitraire sont fondées sur une procédure de segmentation d'image par la couleur.
Les auteurs ont porté une attention particulière aux techniques d'indexation multidimensionnelle, afin d'accomplir le plus efficacement possible des requêtes efficaces dans des espaces d'index de dimension relativement élevée. Ils proposent notamment une approche originale fondée sur un algorithme de regroupement (*clustering*) hiérarchique et dynamique.
L'interface utilisateur du système offre la possibilité de spécifier les requêtes, soit en créant des esquisses colorées et /ou texturées, soit en sélectionnant comme exemple, un objet de la base.
En outre, le système inclut des outils de retour par l'usage (*relevance feedback*) [Porkaew99] permettant d'augmenter la pertinence des résultats de la requête par la création de profils utilisateurs.
Notons également la participation active des auteurs à la normalisation MPEG-7, surtout à ses débuts [Rui97-Iso.a, Rui98-Iso.b].
Mentionnons enfin que cette même équipe de recherche a développé plus récemment (en coopération avec l'Université de Californie - Irvine) un second système d'indexation, nommé WebMars [Ortega00], dédié aux applications de requête par similarité textuelle et visuelle sur Internet.

Le système Informedia [Wactlar00], développé à l'Université de Carnegie Mellon, cible des applications spécifiques liées à la requête par le contenu à propos d'émissions télévisées et de documentaires. Les outils considérés s'adaptent à la spécificité hautement sémantique de ces contenus vidéos, en se focalisant sur des aspects de détection et reconnaissance de la parole et du texte présents dans l'image, pour des procédures de requête/filtrage textuelles. Les descripteurs mis en œuvre sont donc de haut niveau.
Le système inclut toutefois deux composantes purement visuelles, un descripteur de couleurs dominantes [Gong98] et un histogramme global de couleur.
Une procédure de détection et reconnaissance de visage fondée sur la méthode classique des *Eigenfaces* [Turk91] est mise en œuvre à l'aide d'un réseau neuronal.

Les auteurs soulignent l'importance de disposer d'outils de visualisation, présentation et navigation, facilitant l'accès de l'utilisateur aux contenus multimédias à de multiples niveaux d'abstraction. La plate-forme intègre donc plusieurs éléments visuels typiques des présentations multimédias, comme icônes, collages, cartes, en-têtes, bandes de films (*film strips*) et résumés vidéos. Toutefois, une limitation du

système est liée au petit nombre d'outils d'annotation interactive disponibles.

Le système MoCA (*Movie Content Analysis*) [Lienhart96], développé à l'Université de Mannheim (Allemagne), inclut également des procédures de détection et reconnaissance de texte et de visage, conjointement aux outils de structuration à base de détection de plans et scènes. Des approches intéressantes et spécifiques sont intégrées dans le cadre de la composante VisualGREP [Lienhart97], qui propose diverses stratégies pour estimer le degré de la similarité entre séquences vidéos à différentes résolutions et durées (trames, plans, scènes, vidéo globale), en exploitant des techniques d'appariement total ou partiel de chaînes de caractères et de graphes.
Les descripteurs bas-niveau considérés, méthodologiquement plus élaborés que ceux considérés dans les autres systèmes, couvrent tous les attributs visuels habituels. Ainsi, la couleur est-elle représentée par des vecteurs de couleur cohérente (*Color Coherence Vectors* - CCV) [Pass96] dans l'espace de chromaticité uniforme Lab. Les aspects texturaux sont pris en compte par des corrélogrammes d'orientation des contours. Le mouvement est caractérisé par une mesure d'activité exprimant le rapport moyen des modifications de contours sur l'ensemble d'un plan, et la forme par le descripteur MPEG-7 CSS présenté au Chapitre 1, Paragraphe 1.3.5.3.

Le système Netra-V [Deng98], développé à l'Université de Californie à Santa Barbara, propose une représentation totalement orientée objet, enchaînant différents modules de segmentation spatiale et spatio-temporelle et de suivi d'objet. Plus précisément, l'étape de segmentation spatiale par la couleur fournit une initialisation au module de segmentation spatio-temporelle par mouvement paramétrique affine. Le suivi est assuré en exploitant des attributs de forme, texture et mouvement des objets déterminés, qui sont ensuite représentés hiérarchiquement. A chaque niveau hiérarchique, les objets sont indexés par des attributs de couleur (histogrammes de couleur), de texture (les descripteurs MPEG-7 de texture homogène et de parcours rapide de textures), de forme (représentation des contours par descripteurs de Fourier), de mouvement (par paramètres affines) et de localisation par centre de gravité et boîte englobante de l'objet.

Toujours dans le même cadre de la représentation par objet, le système Video-Q [Chang98] de l'Université de Colombie (Etats-Unis) étend aux contenus vidéos les solutions précédemment développées dans le domaine de l'indexation d'images fixes [Smith96]. Les représentations proposées restent toutefois assez élémentaires : histogrammes de couleur dans l'espace LUV, descripteurs de Tamura pour la texture, description grossière de forme par les valeurs principales associées aux deux axes d'inertie de l'objet et trajectoires définies via un ensemble de points. Les problèmes difficiles de requêtes partielles à base de trajectoire ne sont pas abordés, et la mesure de similarité fondée sur la distance $L_2$ entre les points-clé ne garantit pas les propriétés minimales d'invariance.
Le système offre à l'utilisateur une interface graphique lui permettant de spécifier des requêtes par l'esquisse, combinant attributs de texture, de couleur et de trajectoire.

Les systèmes cités ci-dessus donnent un bon aperçu du monde de l'indexation des documents multimédias avant l'élaboration du standard MPEG-7, puisqu'aucun d'entre eux ne pose le problème de

l'interopérabilité, qui implique en outre le stockage et l'échange de plusieurs descriptions associées au même document vidéo.

Cela est d'autant plus important à considérer que l'analyse d'un document audio-visuel fait intervenir simultanément diverses composantes correspondant à différents médias : vidéo, images fixes, objets en mouvement, texte et son. Chaque composante est issue d'un procédé de segmentation (temporelle, spatiale ou spatio-temporelle) et peut-être indexée par des descripteurs adaptés à sa spécificité (Figure 4.1).

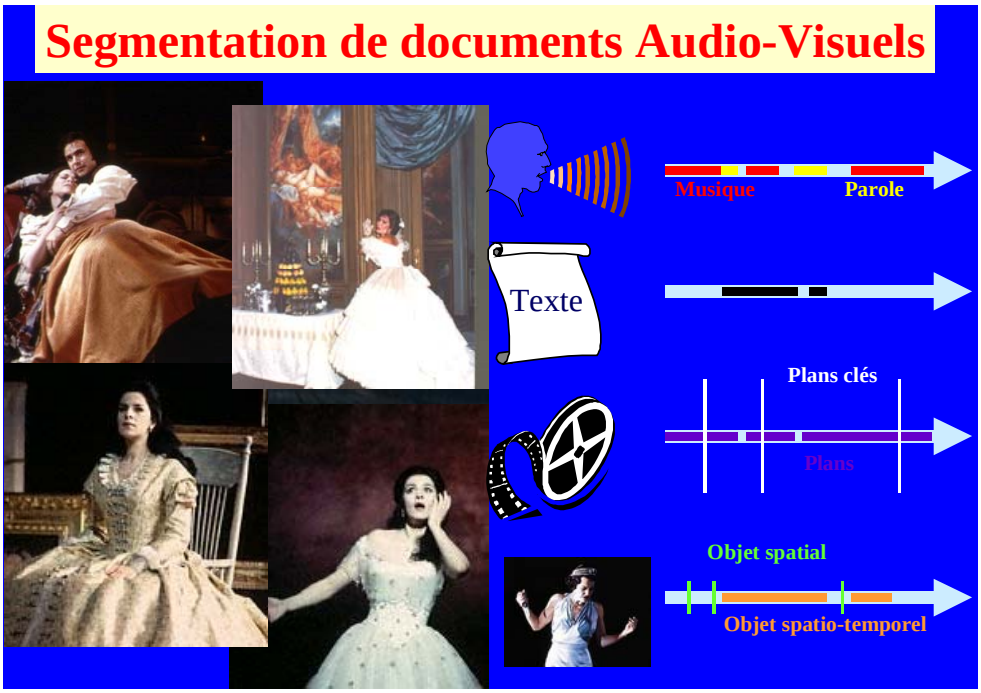



**Figure 4.1.** Différents éléments d'un document audio-visuel
(plans vidéos, images fixes ou en mouvement, texte, son) et segmentations associées.

De la complexité et de la diversité des éléments d'un document audio-visuel découle la nécessité de définir des mécanismes hiérarchiques et génériques de combinaisons de signatures mono et multi-médias. Ces mécanismes devront en particulier permettre la représentation synchronisée de l'information audio-visuelle selon plusieurs niveaux de granularité, correspondant à différents niveaux de hiérarchie.
L'approche adoptée par le futur standard MPEG-7 dans le contexte des descriptions structurales des documents vidéos, que nous décrirons par la suite, permet effectivement de gérer d'une façon unifiée ces différents aspects en levant ce verrou technologique.

## 4.2 L'approche adoptée dans le contexte MPEG-7

Les schémas de description structuraux jouent un rôle central dans l'ensemble des outils MPEG-7. En définissant de manière hiérarchique la structure spatio-temporelle d'un document AV, ils constituent la colonne vertébrale de la description, permettant de gérer d'une manière unifiée les différentes représentations par contenu, tout en intégrant n'importe quel descripteur individuel relatif aux informations de forme, couleur, texture et mouvement.

Pour créer des descriptions récursives, hiérarchiques et adaptées à chaque média, MPEG-7 a adopté une stratégie fondée sur la définition de segments spécifiques à partir d'un segment multimédia générique, appelé *Segment DS*.
Le *Segment DS* est une structure abstraite représentant une "partie" générique d'un document audio-visuel et intégrant les caractéristiques communes à toutes les descriptions. Ses éléments correspond à :

- une composante d'identification définie par *UniqueIdentifier DS*, associant à chaque segment un nom (une chaîne de caractères) permettant de l'identifier de manière unique parmi les différentes autres composantes de la description,
- des composantes temporelles permettant de générer des descriptions synchronisées à l'aide de :
  - *Time DS* : spécifiant globalement la durée de vie d'un segment par son début et sa durée, exprimées en références temporelles (*time stamps*) SMPTE [SMPTE],
  - *TimeMask DS* : précisant de manière plus précise les intervalles temporels non-contigus en définissant les sous-intervalles constitutifs. Dans ce cas, la durée du segment est égale à la longueur du plus petit intervalle de temps couvrant l'ensemble du masque temporel.
- une composante assurant le mécanisme générique de décompositions récursives des segments en sous-segments via la *SegmentDecomposition DS*,
- une composante d'annotation textuelle,
- une composante d'information sur la création du contenu et les droits d'usage,
- une composante définissant l'importance du segment dans une description via la notion de point de vue (*PointofView DS*),
- une composante intégrant des mesures de confiance relatives aux descripteurs utilisés.

Entité abstraite, au sens de la programmation orientée objet, le *Segment DS* ne peut pas être instantié et utilisé tel quel. Il ne prend vie qu'à l'aide du mécanisme d'héritage (Figure 4.2) à travers des segments hérités. Ceux-ci contiennent tous les éléments du Segment DS, auxquels s'ajoutent leurs propres éléments caractéristiques. Chaque type de segment représente le "contenant" (*container*) qui regroupe les descripteurs et les schémas de description spécifiques à chaque type de média.

Parmi tous les segments MPEG-7, nous n'avons considéré que ceux présentés dans le Tableau 4.1.

| | |
|---|---|
| *StillRegion DS* | Une région statique est un ensemble de pixels (connexes ou non) d'une trame de la séquence vidéo. Cette région est obtenue par une procédure de segmentation spatiale. Une image fixe est un cas particulier de *StillRegion DS*. Ce schéma de description intègre l'ensemble des descripteurs d'images fixes en termes de couleur, texture et forme (*cf.* Chapitre 1). |
| *StillRegion3D DS* | Une région statique 3D est un maillage 3D en format VRML. Le seul descripteur associé est celui de spectre de forme 3D (*cf.* Chapitre 2). |
| *Mosaic DS* | La mosaïque définit une image panoramique, construite par superposition de trames successives avec compensation paramétrique de mouvement. Elle permet d'étendre de manière naturelle les descripteurs d'images fixes aux segments vidéos. |
| *VideoSegment DS* | Un segment vidéo est constitué d'un ensemble de trames (temporellement contiguës ou non) d'une séquence vidéo obtenu par segmentation temporelle de la vidéo (par exemple : scène, plan, transition). Une seule trame est un cas particulier de segment vidéo. Le *VideoSegment DS* regroupe tous les descripteurs relatifs aux segments vidéos : attributs de couleur, texture et mouvement global (*cf.* Chapitre 1). En particulier, les descripteurs caractéristiques des images fixes sont intégrés à travers la structure de série temporelle. |
| *AudioSegment DS* | Ce schéma fournit la description du son sur des intervalles temporels respectant certains critères d'homogénéité et intégre les descripteurs audio MPEG-7 [MPEG-7-Audio]. |
| *MovingRegion DS* | Une région en mouvement est définie comme un ensemble de pixels (non nécessairement temporellement ou spatialement connexe) dans un ensemble de trames d'une séquence vidéo, correspondant à une segmentation spatio-temporelle de la vidéo. Les mêmes descripteurs que ceux du *VideoSegment DS* sont ici considérés, avec l'ajout des descripteurs spécifiques de localisation spatio-temporelle. |
| *AudioVisual DS* | Ce schéma de description renvoie à la description de segments temporels (connexes ou non) dans la vidéo mêlant des caractéristiques vidéos et audios. Ce concept permet la fusion des contenus multimédias en une description unifiée et est particulièrement utile pour créer des tables des matières. |

**Tableau 4.1.** Les segments MPEG-7 considérés dans la plate-forme AMIS.

D'autres segments MPEG-7, comme le *Ink DS* (définissant des documents imprimés et incluant des éléments de reconnaissance de l'écriture), le *Multimediasegment DS* (dédié à la représentation des présentations multimédias), le *AudioVisualRegion DS* (similaire au *AudioVisual DS*, mais à un niveau de détails que nous n'avons pas jugé utile de considérer ici), le *ImageText DS* et le *VideoText DS* (dérivés respectivement de *StillRegion DS* et de *MovingRegion* DS par héritage) restent pour l'instant largement au-

delà des objectifs de nos applications et par conséquent n'ont pas été pris en compte dans la version actuelle de la plate-forme AMIS.

En revanche, nous avons intégré des descripteurs en dehors du standard MPEG-7, comme les descripteurs par transformée de Hough, en leur associant des représentations MPEG-7 *ad hoc* définies en DDL MPEG-7.

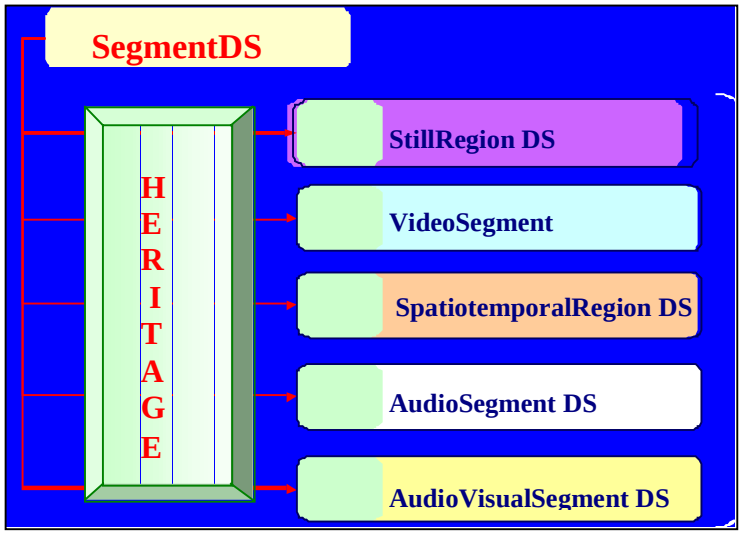



**Figure 4.2.** Illustration du principe d'héritage des segments MPEG-7. Les régions en vert correspondent à des élément communs à tous les SD.

A noter que le même mécanisme d'héritage s'applique également au niveau de la composante *SegmentDecomposition DS*. Les quatre SD de décomposition suivants :

- *Media SegmentDecomposition*,
- *Spatial SegmentDecomposition*,
- *Temporal SegmentDecomposition*,
- *Spatio-temporal SegmentDecomposition*,

sont tout d'abord définis comme structures abstraites, par héritage du *SegmentDecomposition DS*. Ils sont ensuite personnalisés pour chaque type de segment MPEG-7, en appliquant encore une fois le mécanisme d'héritage. Cela permet d'autoriser les seules décompositions pertinentes (en interdisant, par exemple, la décomposition temporelle d'une image fixe représentée par un *StillRegion DS*).

La Figure 4.2 présente un exemple de décomposition faisant intervenir les mécanismes temporels, spatio-temporels et médias.

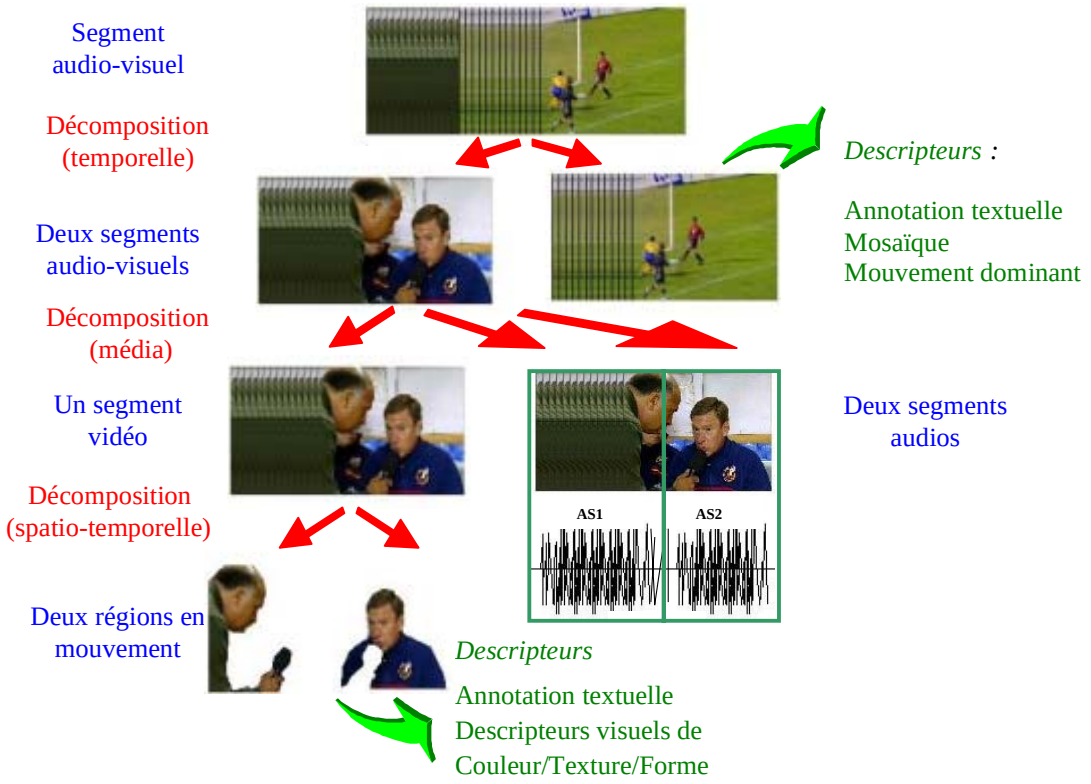



**Figure 4.3.** Exemple de décomposition temporelle et spatio-temporelle sur plusieurs niveaux de hiérarchie.

L'exemple de la Figure 4.3, présente des scènes d'une retransmission sportive, regroupant différents types de segments (audio-visuels, vidéos, audios, région en mouvement) au sein de la structure hiérarchique de description associée.

Trois avantages majeurs de cette approche par décomposition sont à souligner :

- la possibilité de fusionner une véritable information multimédia,
- la possibilité de créer des tables de matières multiples, selon les différents mécanismes de décomposition adoptés,
- la création de description multi-granulaires, un même segment pouvant être représenté aussi bien globalement par ses descripteurs propres, que par l'ensemble des descripteurs associés à ses sous-segments.

Pour supporter une telle approche, le langage de description des données doit satisfaire à des contraintes spécifiques [MPEG-7REQ], comme accepter un mécanisme d'héritage et des descriptions récursives. C'est en partie la raison pour laquelle le DDL MPEG-7 est fondé, modulo quelques extensions spécifiques relativement mineures, sur le langage *XML Schema* [XMLSchema].

Une description d'un document vidéo en DDL MPEG-7 est un fichier en format textuel, faisant intervenir tous les éléments définis ci-dessus et pouvant être parcouru d'une manière interactive et visualisé par des

navigateurs Web, comme par exemple le Microsoft InternetExplore. Des *parseurs* dédiés, comme le Xerces Appache [Xerces], disponibles en source libre et en version C++ et Java, permettent de vérifier l'exactitude de la syntaxe des descriptions.

Toutefois, dans le cadre d'un système d'indexation visant à prendre en compte des documents vidéos de longue durée (entre une et deux heures, typiquement), il est important de disposer d'outils de visualisation, navigation et accès efficaces s'appuyant sur les DS MPEG-7, et reflétant d'une manière intuitive leur nature hiérarchique.
Poursuivons donc, en expliquant comment les différents éléments MPEG-7 MDS ont été pris en compte dans le système AMIS.

### 4.2.1 Le cœur de la plate-forme AMIS

Les fonctionnalités supportées par AMIS visent naturellement à satisfaire les objectifs habituels des systèmes génériques d'indexation notamment en termes de :

- Navigation et accès au contenu au niveau de l'ensemble de la base ou de chaque élément constitutif,
- Requêtes par similarité du contenu, demandant l'intégration d'un moteur de recherche adapté aux descriptions associées,
- Annotation et structuration du contenu,

dans le nouveau contexte de l'interopérabilité MPEG-7.

La plate-forme AMIS s'identifie à un système générique d'indexation normalisée MPEG-7. Elle doit donc offrir un ensemble structuré d'outils, flexible, configurable et extensible, permettant de gérer, d'une manière unifiée, différentes bases de données constituées d'objets multimédias de types aussi différents qu'image fixe, séquence d'images, vidéo, objets 3D, ....

Conceptuellement, une base de données représente une collection d'*objets* multimédias arbitraires. Un objet est ici défini d'une façon générique comme un triplet :

(*média original, média - représentation iconique – description*).

Les composantes *média originale* et *description* sont les éléments indispensables et caractéristiques à toute approche d'indexation. Quant aux représentations iconiques, elles sont définies comme des éléments médias auxiliaires, physiques ou virtuels, fournissant une représentation visuelle symbolique du média original.

Les différents objets actuellement supportés dans AMIS sont décrits dans le Tableau 4.2.

| Média original | Média – représentation iconique | Description |
|---|---|---|
| *Image fixe*<br>(formats ppm, pgm, gif, jpg) | Image fixe | Descripteurs MPEG-7 de couleur, forme, texture, descripteur de Hough2D et descripteur de vecteur de couleurs cohérentes. |
| *Séquence d'images fixes*<br>(trame par trame) | Image fixe de flot optique | Descripteur MPEG-7 de mouvement paramétrique |
| *Maillages 3D*<br>(format VRML) | Image fixe | Descripteur MPEG-7 de spectre de forme et descripteur de Hough 3D optimisé. |
| *Vidéo*<br>(format MPEG-1, MPEG-2, MPEG-4) | Vidéo "virtuelle" (sans dupliquer le média) de petite taille. | Descripteurs et schémas de description MPEG-7 exprimés et stockés en langage de description MPEG-7. |

**Tableau 4.2.** Objets supportés dans la plate-forme AMIS.

Pour les trois premiers types d'objet la puissance de l'approche MPEG-7 se limite à la seule richesse des descripteurs, soit d'images fixes, soit de forme 3D ou encore de mouvement paramétrique. En revanche, elle s'exprime dans toute sa dimension en tirant avantage du caractère hiérarchique et interopérable des descriptions compatibles MPEG-7 dès lors qu'il s'agit d'objets vidéos.

Pour gérer les différents types d'objet et personnaliser les diverses fonctionnalités en fonction de ceux-ci, nous avons adopté une architecture modulaire (Figure 4.4), constituée de trois couches logicielles :

- *Eléments de base*, regroupant les outils communs à tous les objets, et fournissant un modèle générique de gestion de la base qui doit être personnalisé ensuite selon les types de média et les descriptions associées,
- *Gestion de médias*, concernant les aspects liés au type spécifique de chaque objet multimédia et regroupant des outils d'accès au contenu et de visualisation d'objets multimédias,
- *Gestion de descriptions*, regroupant les outils associés à chaque élément de description (descripteur, schéma de description, document en langage de description MPEG-7).

Cette structuration en trois couches permet en outre de gérer aisément d'éventuelles extensions en termes de types d'objet, en n'ajoutant que les modules logiciels relatifs à ceux-ci.

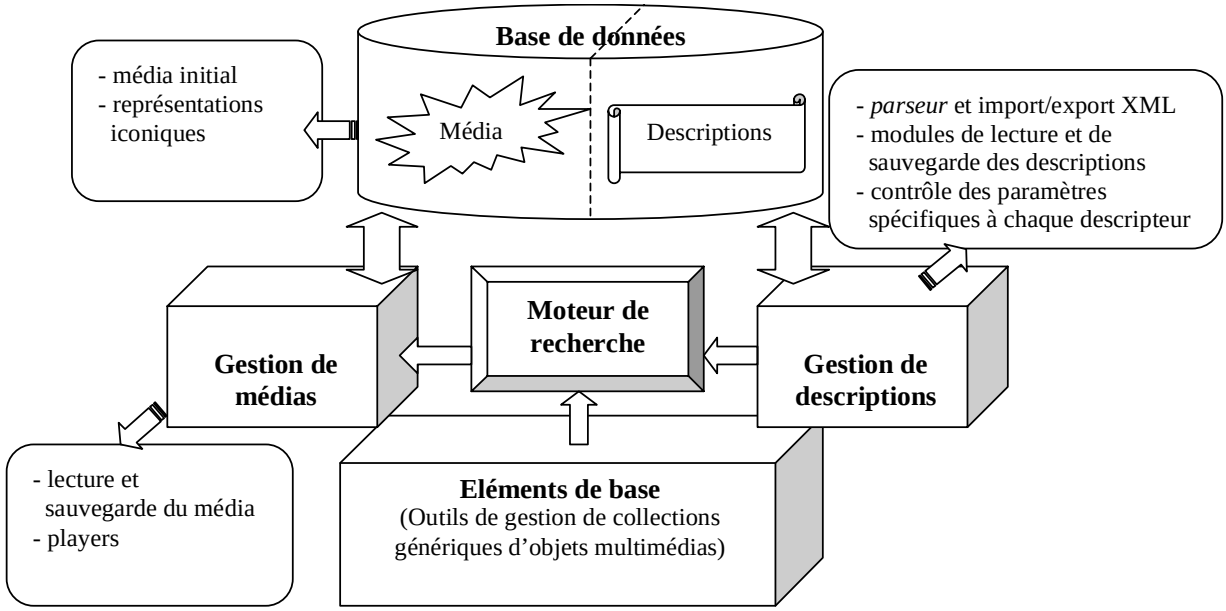



**Figure 4.4.** Les différentes couches d'AMIS.

Au niveau de l'interface avec utilisateur, les différents éléments de chaque couche sont organisés dans différents panneaux abstraits, notés de $P_1$ à $P_5$ et illustrés Figure 4.5.

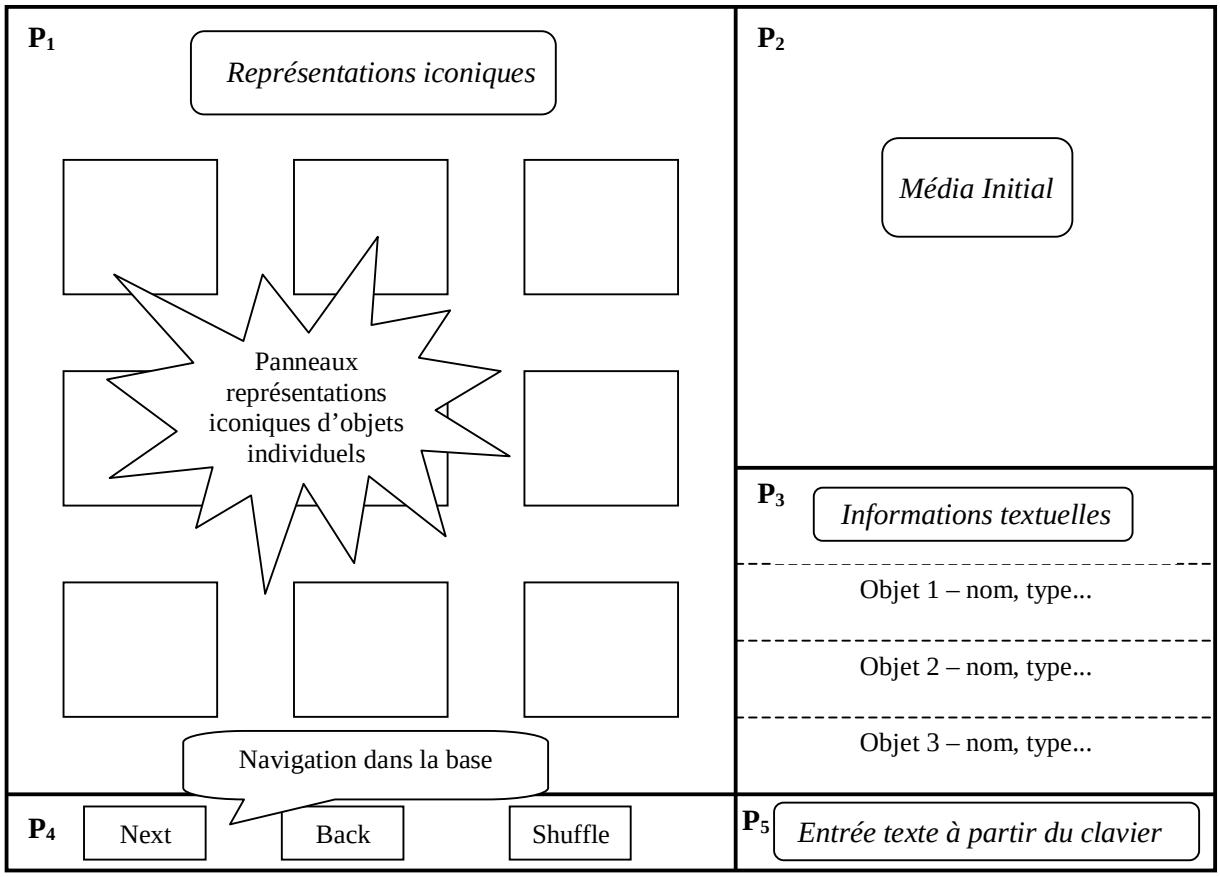



**Figure 4.5.** Les différents panneaux d'AMIS et leur disposition fonctionnelle.

Le panneau $P_1$ est dédié à l'affichage des représentations iconiques. Il présente des versions dégradées du média original, offrant à l'utilisateur un aperçu rapide de son contenu. Le panneau $P_2$ est prévu pour afficher le média original et pour accéder à son contenu (sélection de régions, vidéo cliquable, manipulation de maillages 3D). Le panneau $P_3$ présente les éléments textuels relatifs au média (nom de fichier, type...). Pour permettre l'accès par mot clé aux éléments de la base, le panneau $P_4$ permet de saisir du texte et de se positionner sur l'élément de la base trouvé. Enfin, le panneau $P_5$ inclut les fonctionnalités de navigation dans la base, à l'aide de trois boutons, page suivante, précédente et aléatoire. Une page est ici considérée comme un ensemble d'objets en nombre égal à celui des représentations iconiques affichées dans le panneau $P_1$ et contrôlable à partir d'une fenêtre de dialogue spécifique.
L'inclusion dans le système d'un sous-ensemble de tous ces panneaux permet de définir différents modes de fonctionnement, plus ou moins dédiés.

Pour chaque type d'objet, des panneaux caractéristiques sont ensuite définis à partir de ces panneaux génériques, par héritage et ajout d'éléments propres aux types d'objet considérés. Cette approche donne au système un aspect proteïforme, son apparence s'adaptant automatiquement au type d'objet de la base de données considérée.
La Figure 4.2 présente un *snapshot* de l'interface, en mode moteur de recherche, pour la base de maillages 3D MPEG-7.

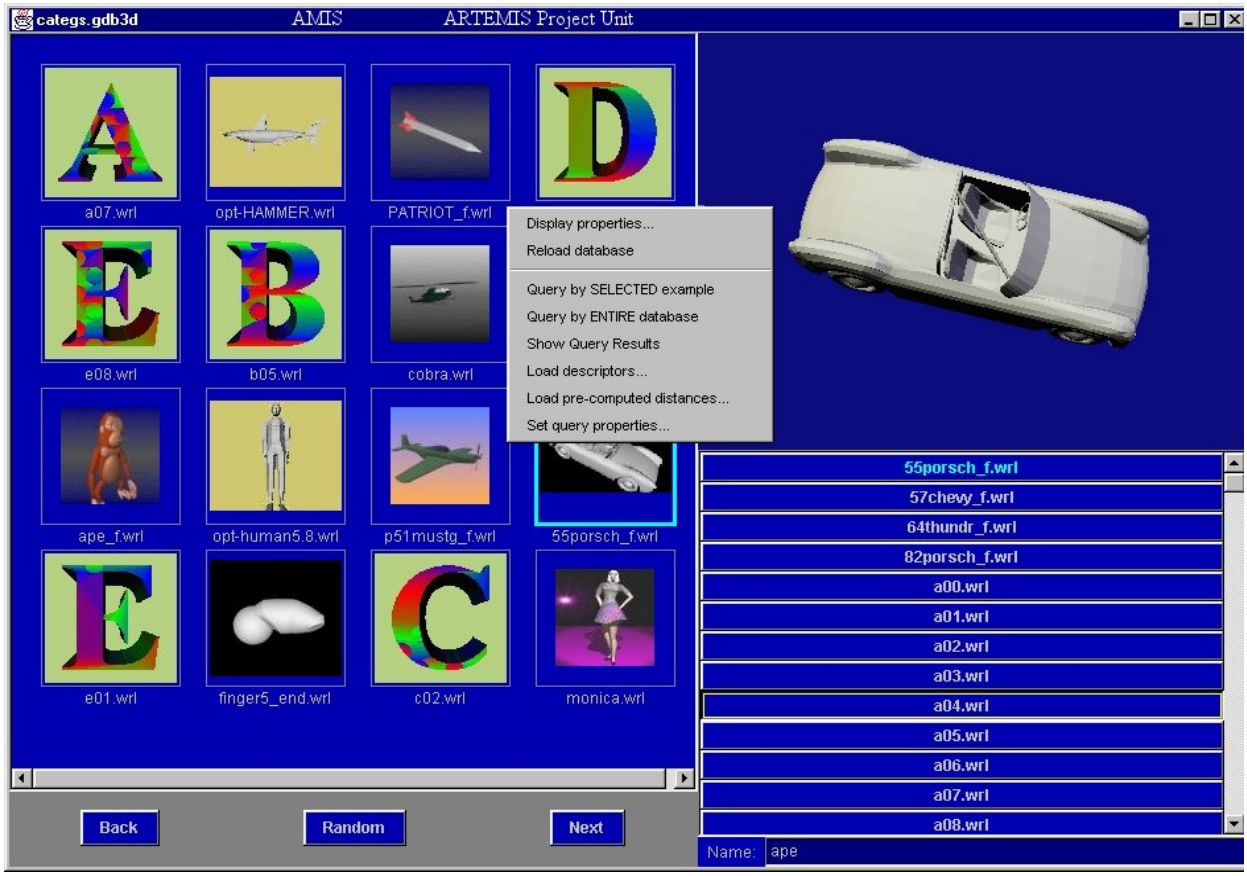



**Figure 4.6.** Moteur de recherche de maillages 3D et gestion des commandes relatives aux descripteurs et aux requêtes (fenêtre de type *pop-up* activée par clic du bouton droit de la souris).

Les représentations iconiques considérées ici sont des imagettes au format JPEG correspondant aux projections 2D des maillages 3D. La Figure 4.2 montre également la fenêtre de type *pop-up*, activée par clic du bouton droit de la souris qui permet, d'une manière personnalisée selon le type de panneau considéré, la gestion des différentes commandes : chargement des descripteurs, lancement des requêtes, contrôle des paramètres d'affichage, lancement des boîtes de dialogue pour choisir parmi les différentes mesures de similarité et ajuster les paramètres associés...

La Figure 4.7 illustre l'apparence d'AMIS, toujours en mode moteur de recherche, mais cette fois pour des séquences de gestuelles indexées par descripteur paramétrique de mouvement.

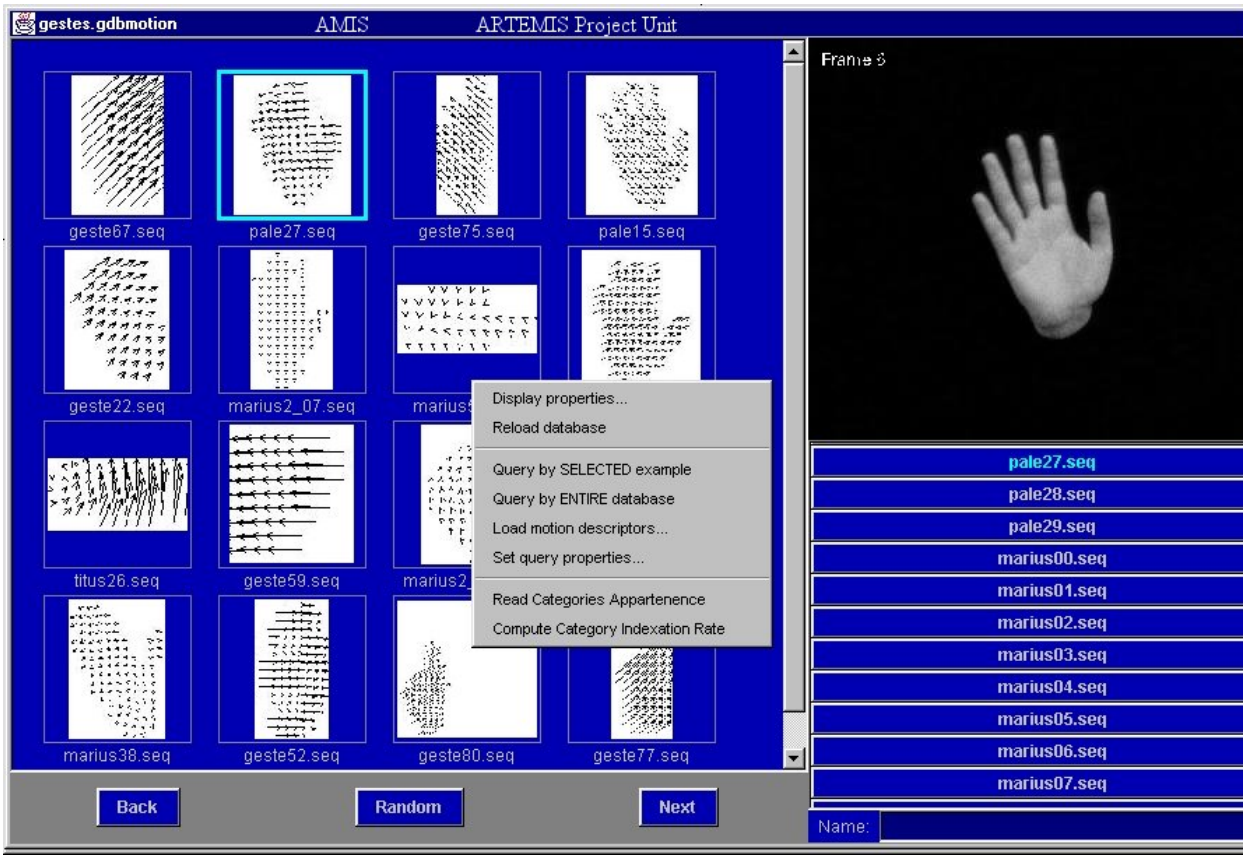



**Figure 4.7.** Moteur de recherche par similarité de mouvement de séquences vidéos de geste et fenêtre de dialogue pour le choix de la mesure de similarité et des paramètres de requête.

Les icônes représentent ici des images du flot optique associé à chaque modèle paramétrique, offrant un aperçu du contenu dynamique des séquences. La fenêtre de dialogue (au centre) permet de sélectionner le type de mesure de similarité et les paramètres associés, comme la fonction distance et le facteur de pondération $k$ (*cf.* Chapitre 2, Paragraphe 2.5.4).

Enfin, en ce qui concerne les documents audio-visuels, la stratégie de présentation et visualisation a été légèrement modifiée, afin de prendre en compte efficacement le volume important et le caractère temporel de ces données.


**Figure 4.8.** Structuration des panneaux AMIS pour le traitement des documents vidéos (corpus AGIR).

Le panneau des représentations iconiques est ici décomposé en une série de panneaux horizontaux affichés les uns en dessous des autres, chaque panneau horizontal comportant les éléments iconiques associés à une seule vidéo. Cette structuration supplémentaire permet notamment de prendre en compte le caractère temporel des données.
Quant aux représentations iconiques elles-mêmes, elles intègrent des *players* permettant la visualisation dégradée en taille et éventuellement en nombre de trames par seconde des différents segments MPEG-7.

Le type de segment visualisé à un instant donné peut être sélectionné à partir de la barre de menu présente dans l'en-tête de chaque panneau vidéo.
Chaque segment est visualisé sur une période définie par les références temporelles incluses dans les SD MPEG-7 et directement à partir du média original, sans avoir besoin de dupliquer les médias pour créer des représentations iconiques. Ce principe de représentation iconique par vidéos virtuelles, permettant l'accès aléatoire et concomitant par différents *players* au média original, représente un des grands avantages du système, puisqu'avec l'avènement des applications MPEG-7 interopérables, la quantité de descriptions multimédias est susceptible d'augmenter d'une façon spectaculaire. Il est alors important de disposer

d'outils supportant plusieurs descriptions du même contenu, sans avoir à créer et à stocker à chaque fois les représentations iconiques associées.

Des outils dédiés permettent de naviguer dans les vidéos selon la structure hiérarchique de la description associée. L'utilisateur peut donc choisir le niveau hiérarchique et descendre ou remonter dans la structure (Figure 4.9) et cela selon les différents types de segments sélectionnés.


**Figure 4.9.** Navigation dans la structure hiérarchique de description (corpus AGIR).

En outre, notons que les requêtes par similarité peuvent porter aussi bien sur toute la base que sur la seule vidéo considérée. Cela permet à l'utilisateur de réduire le temps de réponse aux requêtes, lorsqu'il recherche des éléments à l'intérieur d'une même vidéo, et de bénéficier d'une présentation plus structurée des résultats de la requête.

La visualisation du média original s'effectue dans le panneau $P_2$, les commandes associées permettant notamment d'initialiser et de lancer les algorithmes de segmentation spatiale ou spatio-temporelle ou de lancer l'application de vidéo cliquable (Paragraphe 4.3.1).
Lorsqu'il s'agit uniquement de la navigation et de la visualisation, un mode de fonctionnement simplifié, avec un ensemble réduit de fonctionnalités est proposé (Figure 4.10).

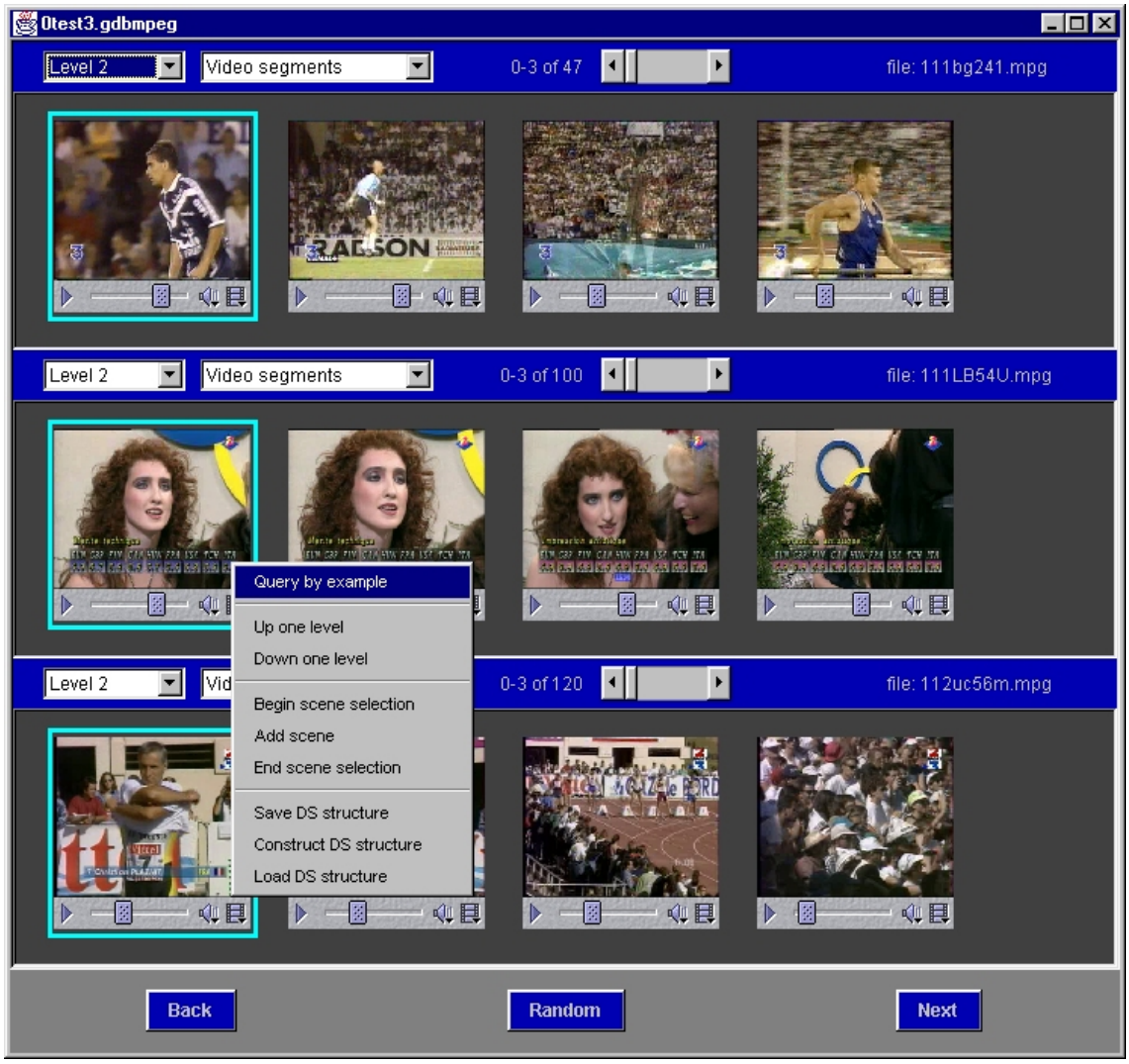



**Figure 4.10.** Mode simplifié de fonctionnement :
les panneaux $P_2$, $P_3$ et $P_5$ ont été éliminés (corpus AGIR).

Les outils d'annotation actuellement intégrés dans le système sont liés à l'application d'archivage de vidéos indexées et de reconnaissances des gestes dans la langue des signes française (*cf.* Paragraphes 4.3.1 et 4.3.2) et regroupent des procédures de segmentation temporelle et spatio-temporelle et tous les descripteurs bas-niveau MPEG-7.

Pour la segmentation temporelle[Corridoni98, Nagasaka92, Hampapur94, Zabih95, Bouthemy99, Joly94], nous avons adopté et développé l'algorithme à base d'histogrammes de couleur décrit dans [Corridoni95], qui permet, avec des performances raisonnables, de détecter aussi bien des transitions brusques (*cuts*) que progressives (volets, dissolutions, ...). Toutefois, comme aucune procédure de segmentation de vidéos quelconques ne garantit des résultats corrects à 100%, le système permet d'éditer et de corriger manuellement la segmentation obtenue. Parce que le regroupement des plans en scènes, est une étape difficilement réalisable de manière automatique dans le cas de vidéos génériques, en raison de leur caractère hautement sémantique, le même outil d'édition offre toute l'interactivité nécessaire pour construire des

tables des matières hiérarchiques. Enfin, ajoutons que la procédure de segmentation en régions de mouvement cohérent décrite Chapitre 2 est intégrée dans le système.

Dans le cas de la segmentation spatiale et spatio-temporelle, il est souvent utile de disposer d'outils semi-automatiques permettant une initialisation ou un apprentissage des principales caractéristiques de l'objet, rapide et aisément supervisé par l'utilisateur. Des outils ont été intégrés pour sélectionner des régions (boîtes rectangulaires, polygones, formes libres) de la scène. Une fois les régions d'intérêt acquises, elles sont ensuite utilisées :

- soit pour effectuer un apprentissage de leur couleur (par simples histogrammes ou par mélange de gaussiennes [Hammoud00]), servant d'initialisation à un algorithme rapide de segmentation par la couleur,
- soit comme initialisation spatiale du masque de l'objet à segmenter, dans le cadre d'un algorithme markovien de suivi d'objet faisant coopérer des informations de couleur, mouvement et contour.

Les objets obtenus sont présentés à l'utilisateur pour validation et intégration dans les descriptions vidéos soit comme régions fixes, soit comme régions en mouvement

La seconde famille d'outils d'annotation concerne aussi bien les descripteurs bas-niveau MPEG-7 de couleur, texture, forme et mouvement (*cf.* Chapitre 1), repris à partir du logiciel de référence MPEG-7 et intégrés via les procédures spécifiques JNI (*Java Native Interface*). Par souci d'homogénéité et de simplicité, le moteur d'extraction utilise la même interface de contrôle de paramètres que celle du moteur de recherche, avec la possibilité supplémentaire de lancer l'extraction en *batch*, à partir d'un certain niveau hiérarchique et pour tous les objets dont le type est spécifié dans la boîte de dialogue associée au moteur d'extraction.

Cette approche permet de créer d'une façon rapide et efficace des descriptions de vidéos, hiérarchiques, riches et interopérables en raison de leur compatibilité MPEG-7.

### 4.2.2 Les outils de développement

Le système AMIS a été intégralement développé sous une plate-forme Java (*Java Development Kit 1.3*) [JDK], en raison des nombreuses facilités offertes en termes de programmation et support multimédia. Parmi celles-ci, citons :

- le support de différents types de médias (image, vidéo, son) et de formats de données compressées (GIF, JPEG, MPEG) ou non, notamment à travers les outils spécifiques de la bibliothèque *Java Media Framework* (JMF) [JMF] pour les données vidéos et de la bibliothèque Java 3d [Java3D] pour les données 3D,
- le support de langages de description de données et d'un *parseur XML Schema* [Xerces] en source libre et sous une licence de type *Royalty Free*,

- le fait de disposer d'un ensemble d'outils graphiques riche et facile à utiliser, permettant de créer rapidement des interfaces personnalisées,
- le support des applications *Internet*, via les *applets*,
- les méthodes de programmation orientée objet (POO) et le mécanisme de *multithreading*,
- la possibilité d'intégrer du code C++ à travers les outils de *Java Native Interface* (JNI),
- l'environnement multi-plateforme (*Windows, Solaris, Linux*).

Les diverses fonctionnalités décrites précédemment ainsi que l'approche logicielle détaillée Paragraphe 4.2.1 ont été intégrées dans la plate-forme AMIS dont on propose maintenant de démontrer l'intérêt pour des applications d'indexation, qu'elles soient génériques comme l'archivage de documents vidéos ou très spécifiques comme l'application d'indexation de la langue des signes pour déficients auditifs.

## 4.3 Les applications MPEG-7 développées

La plate-forme AMIS a été validée dans le cadre de deux applications différentes. La première, développée dans le cadre du projet RNRT [RNRT] AGIR [AGIR] concerne l'indexation générique de bases vidéos et la vidéo cliquable. La seconde est liée à la reconnaissance de gestes en langue des signes française, dans le cadre de la représentation et la requête par le contenu visuel dynamique. Cette application s'inscrit dans le cadre des projets européens Socrates PUZZLE 1 et PUZZLE 2.

### 4.3.1 Indexation d'archives vidéos et vidéo cliquable

Selectionné au premier appel du RNRT (Réseau National de la Recherche en Télécommunications), le projet AGIR (*Architecture Globale pour l'Indexation et la Recherche de documents multimédias*), regroupant partenaires du monde industriel (Communications and Systems, Arts Vidéo Interactive, Mémodata), institutionnel (Institut National de l'Audiovisuel - INA) et académique (INRIA, INT, Université Joseph Fourier, Laboratoire Image et Parole de l'Université Paris VI), a démarré officiellement le 01/01/2000, en se proposant comme objectif de :

*"développer des technologies et des outils nécessaires pour mettre en oeuvre une architecture globale pour l'indexation et la recherche par le contenu de données multimédia, conforme aux exigences exprimées dans le contexte de la normalisation internationale"*.

Pour traiter de cet objectif, le projet AGIR a été structuré en neuf sous-projets (SP), qui sont les suivants :

- SP1 - gestion du projet,
- SP2 - extraction de descripteurs d'image fixe (relatifs aux attributs de couleur, texture et forme),
- SP3 - extraction de descripteurs vidéos (mouvement affine, classification du type de mouvement global, classifications des transitions) et segmentation temporelle en plans, détection et suivi des régions en mouvement,
- SP4 - extraction des signatures texte,

- SP5 - segmentation de la bande sonore et clasification en parole, musique, et sons complexes,
- SP6 - extraction des signatures monomédias composites et multimédias,
- SP7 - élaboration du langage de description, de schémas de description et leur instanciation,
- SP8 - développement du moteur de recherche,
- SP9 - architecture globale et applications.

Le schéma générique du projet AGIR est présenté Figure 4.11. Notons que le SP6, dont INT-ARTEMIS était responsable, joue un rôle tout à fait particulier puisqu'il fédère, en les structurant par combinaison, fusion, coopération, les résultats des sous-projets d'extraction de signatures isolées sur des supports monomédias. Il s'interface, via la phase d'instanciation des schémas de description, au moteur d'indexation et de recherche vidéo. En outre, pour assurer la généricité des aspects de composition, fusion et coopération des signatures, il prend en compte les spécifications du DDL. INT-ARTEMIS a apporté son expertise de la normalisation MPEG-7 au niveau du SP7 où il était contributeur principal.

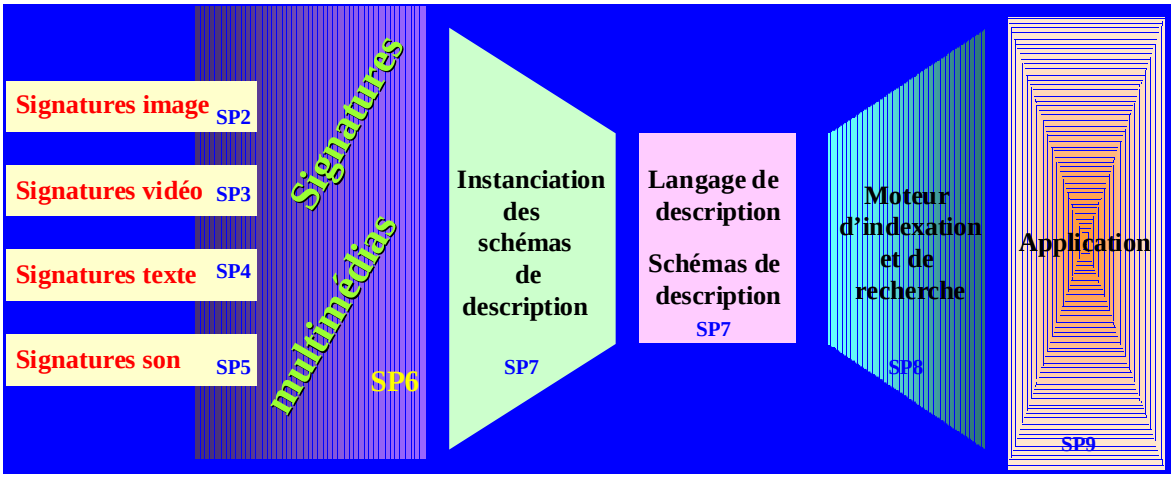



**Figure 4.11.** Schéma synoptique du projet AGIR et position du SP6..

Un corpus de travail a été constitué par l'INA et distribué à tous les partenaires du Consortium AGIR. Il comporte une dizaine d'heures de vidéos thématiques au format MPEG-1, correspondant à différentes retransmissions sportives.

Le projet AGIR, conduit par l'INA, s'est articulé autour de deux orientations technologiques, l'une propriétaire, l'autre standardisée MPEG-7 :

- La première, proposée par l'INA, consiste en des descriptions "linéaires", exprimées à l'aide du langage de description AEDI-4 (à base de XML), élaboré par l'INA, démontre la faisabilité d'un système d'indexation d'archives vidéos. Toutefois, elle ne supporte pas de façon naturelle des mécanismes d'héritage et de récursivité, qui en limite l'efficacité fonctionnelle et l'interopérabilité.
- La seconde, fondée sur MPEG-7 et développée par INT-ARTEMIS, exprime les différents descripteurs AGIR et les descriptions des documents audio-visuels en *XML Schema*. Fondée sur la plate-forme AMIS, elle a été validée par les partenaires du projet et présentée lors de l'audit

d'AGIR le 26/09/01 devant les représentants du Ministère de la Recherche et de l'Industrie. La plate-forme AMIS a pleinement démontré la supériorité de l'approche MPEG-7, en termes de facilités de navigation, avantages des descriptions granulaires et hiérarchiques et capacité de stockage des descriptions. C'est aujourd'hui cette orientation unifiée et normalisée dont nous avons prouvé le caractère opérationnel que le Consortium souhaite adopter.

Comme il s'agissait ici de fédérer segmentations et descripteurs différents, fournis de manière indépendante par les autres sous-projets, nous avons développé des outils spécifiques pour :

1. Créer l'arbre de description à partir des divers segmentations de type vidéo et audio fournis par les divers sous-projets d'analyse,
2. Insérer des descripteurs spécifiques à chaque type de média. Ici, on suppose que des listes distinctes de descripteurs, spécifiés en *XML Schema*, sont disponibles comme entrées dans le système. Pour en assurer la synchronisation, un champ complémentaire est inséré dans chaque descripteur individuel, spécifiant soit le numéro de la trame considérée, dans le cas des descripteurs d'images fixes, soit un intervalle temporel, pour les autres descripteurs.

Les Figures (Figure 4.12, Figure 4.13, Figure 4.14) présentent quelques exemples de requêtes par similarité, selon différents types de segment, à divers niveaux hiérarchiques et exploitant différents descripteurs de mouvements ou de couleur.

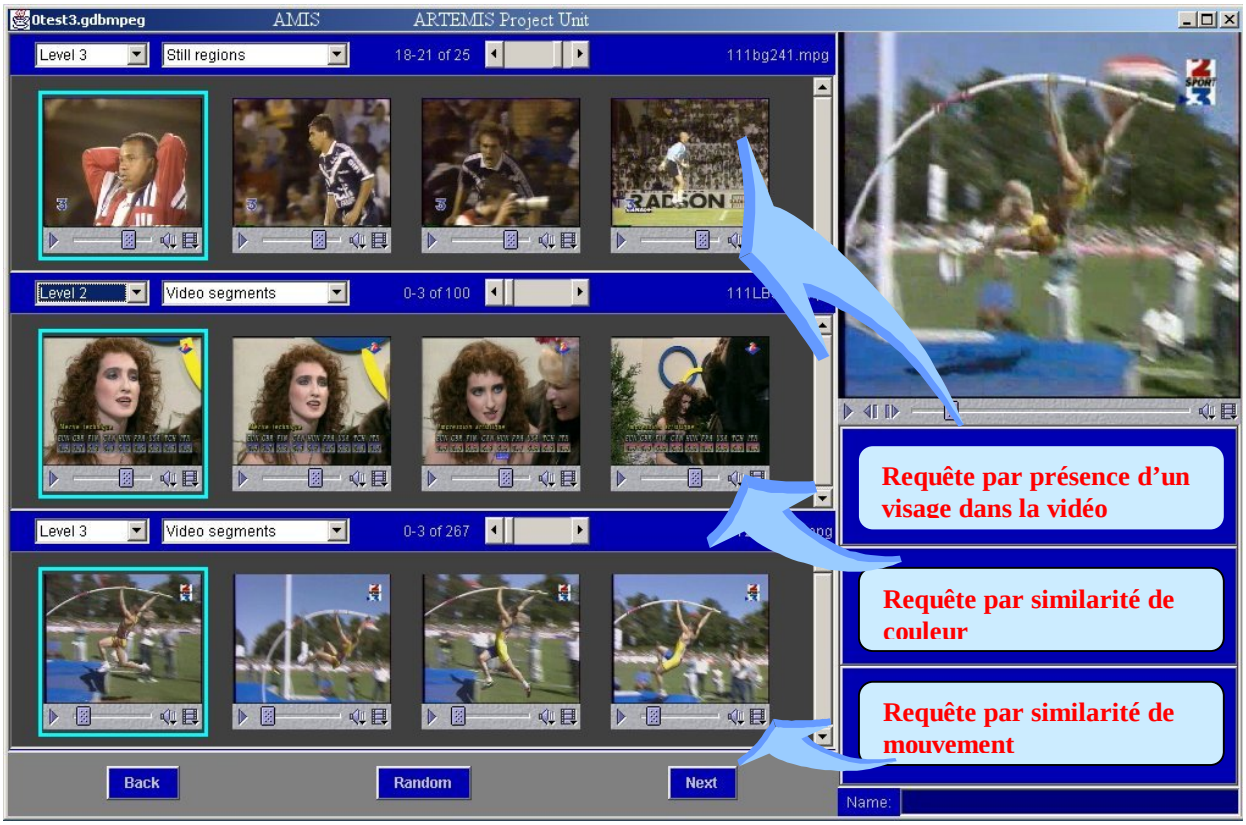



**Figure 4.12.** Exemples de requêtes.

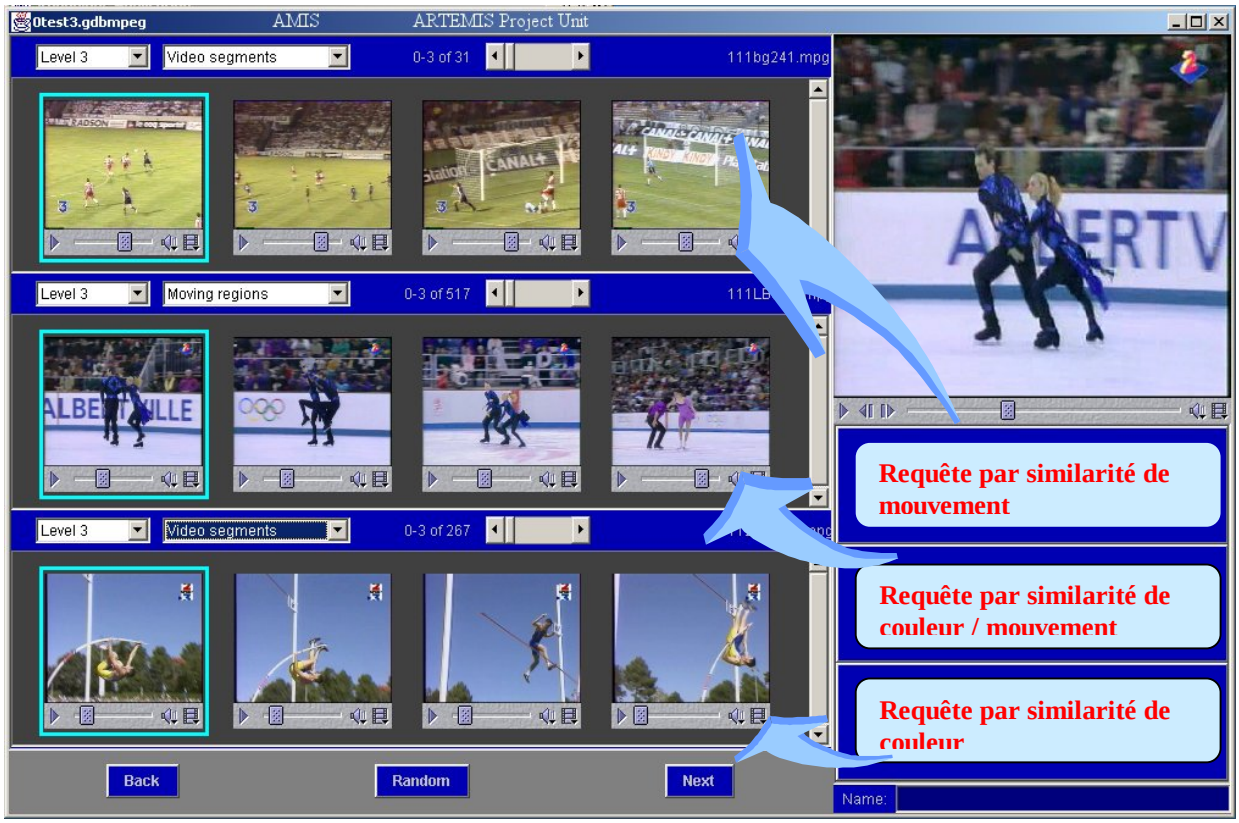



**Figure 4.13.** Exemples de requêtes.


**Figure 4.14.** Exemples de requêtes.

Afin de profiter pleinement des fonctionnalités MPEG-7, notamment en termes de descriptions orientées objet vidéo, nous avons intégré et développé une application de vidéo clicable, en nous inspirant des systèmes déjà existants expérimentés mais propriétaires [ArtsVidéo].

L'application de vidéo cliquable est lancée en cliquant sur l'objet d'intérêt (les patineurs illustrés Figure 4.15 et Figure 4.16). Une boîte de dialogue apparaît, fournissant divers éléments d'information textuelle avec des hyper-liens actifs, permettant d'accéder directement aux sites Internet relatifs à ces "objets". La plate-forme permet également de selectionner un objet dans l'image et de lancer de requêtes par similarité exploitant les descripteurs associés.


**Figure 4.15.** Exemple dans l'application de vidéo cliquable.

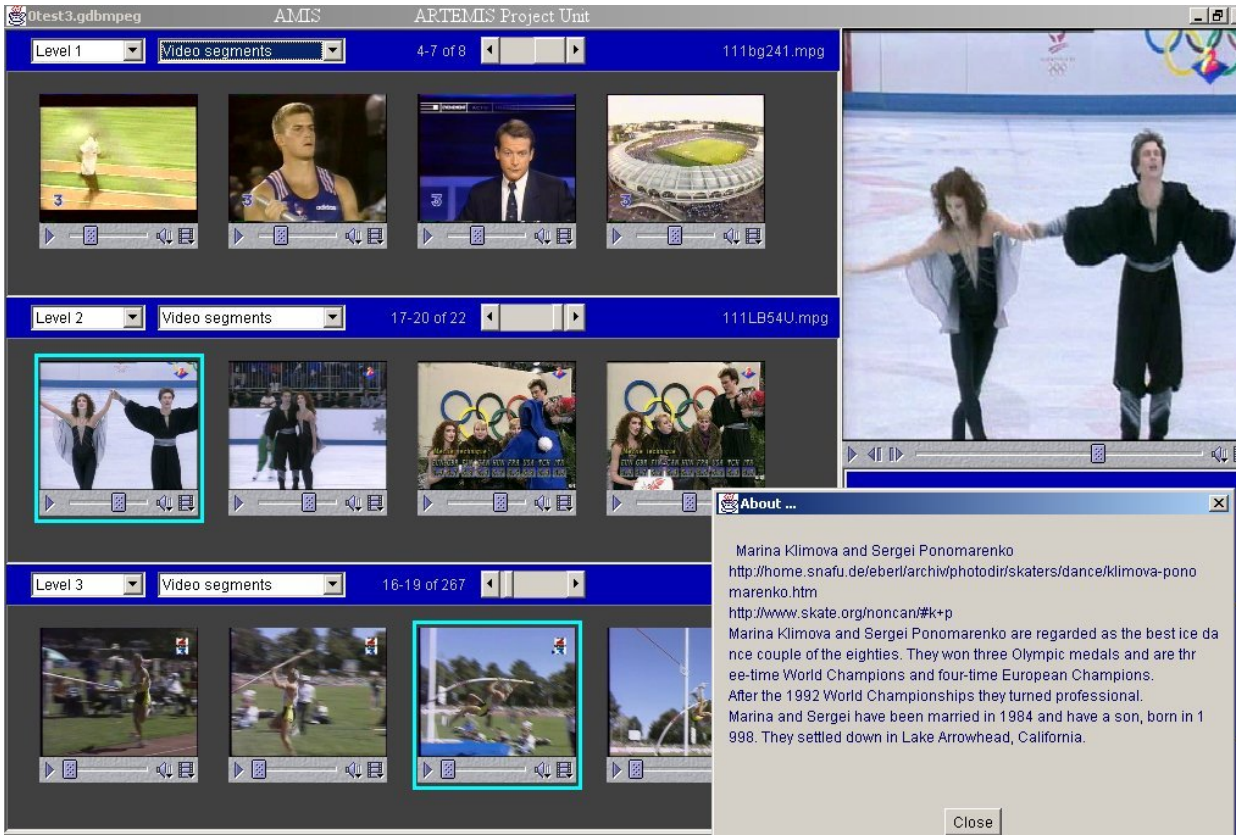



**Figure 4.16.** Exemple d'application de vidéo cliquable.

L'ensemble logiciel ainsi réalisé démontre pour la première fois en grandeur réelle, pour les applications d'indexation multimédia, comme l'archivage vidéo et la vidéo cliquable, le caractère effectivement opérationnel des schémas de description audio-visuels génériques, normalisés MPEG-7 et ses avantages objectifs en termes d'interopérabilité et réutilisation des descriptions hiérarchiques et multigranulaires.

## 4.3.2 Indexation en langue des signes française

Ce paragraphe traite d'une application spécifique, liée à la reconnaissance des gestes en langue des signes française, dans le cadre de la représentation et la requête par le contenu visuel dynamique.

La langue des signes est le moyen de communication le plus pratiqué par les personnes souffrant de déficiences auditives. Habituellement, une langue des signes a un contenu lexical d'environ 6000 gestes, exprimant un large éventail de concepts, actions, émotions... Un mot peut être exprimé aussi bien par un unique geste qu'en mode dactylographique épelé (*finger spelling*), à l'aide d'un alphabet associant un geste à chaque lettre. Ce dernier mode de communication est surtout utilisé pour exprimer des noms propres, ou des concepts peu usuels.

Comme il n'existe pas un langage parlé universel, il n'y a pas non plus une unique langue des signes, chaque langue des signes étant fortement attachée aux caractéristiques et spécificités culturelles de chaque peuple.

De nombreuses recherches ont tenté d'améliorer les possibilités de communication entre déficients auditifs, par exemple par transmission de conversations en langue des signes sur des réseaux téléphoniques [Mozelle98], par la création de dictionnaires plus ou moins interactifs de langues des signes [CyberSign], par la mise en place d'applications éducatives [DILS] pour les jeunes sourds [Seamless] ou encore à travers l'offre de services commerciaux en langue des signes [Visicast].

La majorité des outils existants s'appuie sur des annotations textuelles et hérite donc de leurs limitations bien connues. C'est pourquoi de nouvelles approches exploitant directement les représentations par le contenu se font jour. Il s'agit principalement :

- d'approches statistiques, généralement inspirées des méthodes de reconnaissance de la parole et fondées sur des chaînes de Markov cachées [Schlenzig94, Starner95, Starner98] pour prendre en compte le contenu dynamique d'un geste,
- d'approches déterministes, utilisant des modèles de main 3D plus ou moins complexes [Rehg93, Dorner96] et s'appuyant sur le suivi 3D de la main à partir d'images 2D, en vision mono- ou multi-caméra.

D'autres techniques sont considérées également dans le domaine connexe de l'interaction homme-machine [Freeman94, Freeman95, Quek94, Quek95, Kahn96], dans le cadre d'applications spécifiques, limitant en général drastiquement le nombre des différents gestes à reconnaître.

Notre recherche sur la langue des signes s'inscrit dans le contexte de la vision monoscopique et ne fait appel à aucun capteur physique, marqueur, ou autre dispositif d'acquisition très spécifique. L'enjeu consiste à identifier et extraire à partir des séquences vidéos 2D des primitives pertinentes permettant une reconnaissance efficace des gestes.

Nos démarches ont été grandement facilitées grâce au concours des experts en langue des signes de l'INJS (Institut National des Jeunes Sourds) et du CEPSAS (Centre Européen de Promotion Sociale des Adultes Sourds), qui nous ont aidés à réaliser les corpus de test pour les expérimentations et à spécifier les attributs susceptibles de discriminer les signes.

#### 4.3.2.1 Corpus de test et création des prototypes "gestuels" naturels et synthétiques

Les séquences vidéos naturelles ont été acquises avec l'aide du Centre National d'Etudes en Télécommunications (CNET) et en coopération avec des experts de l'INJS. Deux corpus différents ont été réalisés :

- Le corpus "Lettres", correspondant aux 28 lettres de l'alphabet (Figure 4.17.a) (environ 4500 trames), acquis en cadrage large (caméra fixée à proximité du signeur, de telle sorte que la main occupe une zone importante de l'image),

- Le corpus "Mots" (Figure 4.17.b), constitué des noms de stations du métro parisien (10 minutes de vidéos, à une cadence de 20 trames par seconde), acquis en cadrage large (la caméra est suffisamment éloignée pour que le signeur soit inclus dans le champ), ce qui correspond de manière plus fidèle aux vidéos que les malentendants sont habitués à voir.

a. Séquence du corpus “Lettres” (cadrage étroit).

b. Séquence “Kleber” du corpus “Mots”(cadrage large).

**Figure 4.17.** Séquences des deux corpus de test avec différentes configurations de la main.

Toutes les séquences ont été numérisées au format QCIF. Les signes consécutifs ont été délimités par l’introduction d’une configuration neutre (paume ouverte, doigts tendus, largement écartés et pointant vers le haut).

Nos discussions avec les experts en langue des signes nous ont permis de tirer profit d’une taxonomie des gestes déjà existante [DILS] et utilisée habituellement pour l'enseignement des langues des signes. Les six principaux attributs sur lesquels repose cette classification exploitent la structure sémantico-syntaxique d'un geste. Il s’agit :

- du nombre de mains utilisées pour signer,
- des positions relatives des mains par rapport à la tête au buste du signeur,
- de la direction des doigts (vers le haut, vers le bas, vers la droite, vers la gauche, vers l'avant, vers l'arrière),
- de l’orientation de la paume (vers le haut, vers le bas, vers droite, vers la gauche, vers l'avant, vers l'arrière),
- de l’action de la main (vers le haut, vers le bas, vers la droite, vers la gauche, vers l'avant, vers l'arrière, rotation, vibration, ondulation),
- de la configuration de la main, *i.e.* la position relative des doigts [DILS, CyberSign]

Parmi ces différents attributs, les plus importants, selon les experts, sont les deux derniers, et le plus discriminant est celui de configuration. Par conséquent, notre objectif premier a consisté en la représentation et la reconnaissance efficaces et fiables de la configuration de la main.

Le suivi efficace de ces caractéristiques dans une séquence gestuelle représente un problème difficile, en raison des nombreuses difficultés à surmonter, liées à la projection en 2D d'un objet 3D articulé et fortement déformable comme la main. Il s'agit typiquement d'inférer une information 3D à partir de données 2D incomplètes, avec des ambiguitées dues aux occultations de tout type, aux déformations globales et locales et aux effets d'ombrage.

Afin de surmonter ces obstacles, nous exploitons la nature spécifiquement dynamique d'un signe, caractérisée par des transitions entre configurations temporellement stables. Le principe de l'approche développée de façon originale consiste à segmenter la vidéo en segments temporels correspondant à une configuration stable de la main (Figure 4.18).

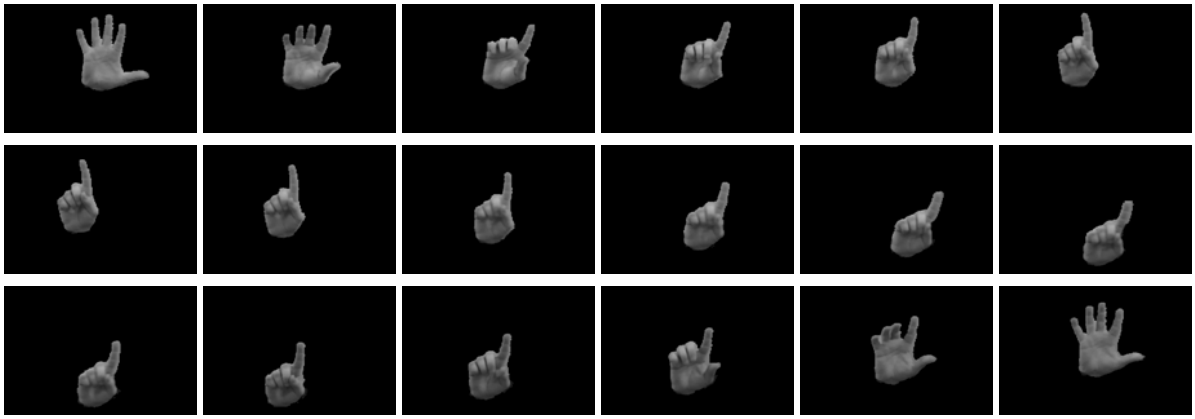


**Figure 4.18.** Lettre 28 ("Z") :les transitions entre les configurations (au début et à la fin du geste) et l'intervalle de configuration stable.

Nous avons alors créé deux types de configurations prototypes :

- les prototypes naturels (Figure 4.19) qui sont extraits directement à partir des séquences vidéos naturelles, et qui sont définis selon le principe suivant. Chaque lettre de l'alphabet correspond à un signe unique et donc à une seule configuration stable de main. Par conséquent, une unique configuration 2D a été choisie comme prototype associé au geste correspondant.
- les prototypes synthétiques (Figure 4.20), représentés par un maillage 3D de main droite dans la configuration considérée, et définis comme les images des projections dans la scène d'un modèle de main articulé et maillé [Preda99]. Les paramètres 3D correspondant aux différentes configurations sont ici totalement connus et exprimés de manière standardisée à l'aide des paramètres d'animation (*Body Animation Parametres* –BAP) MPEG-4.

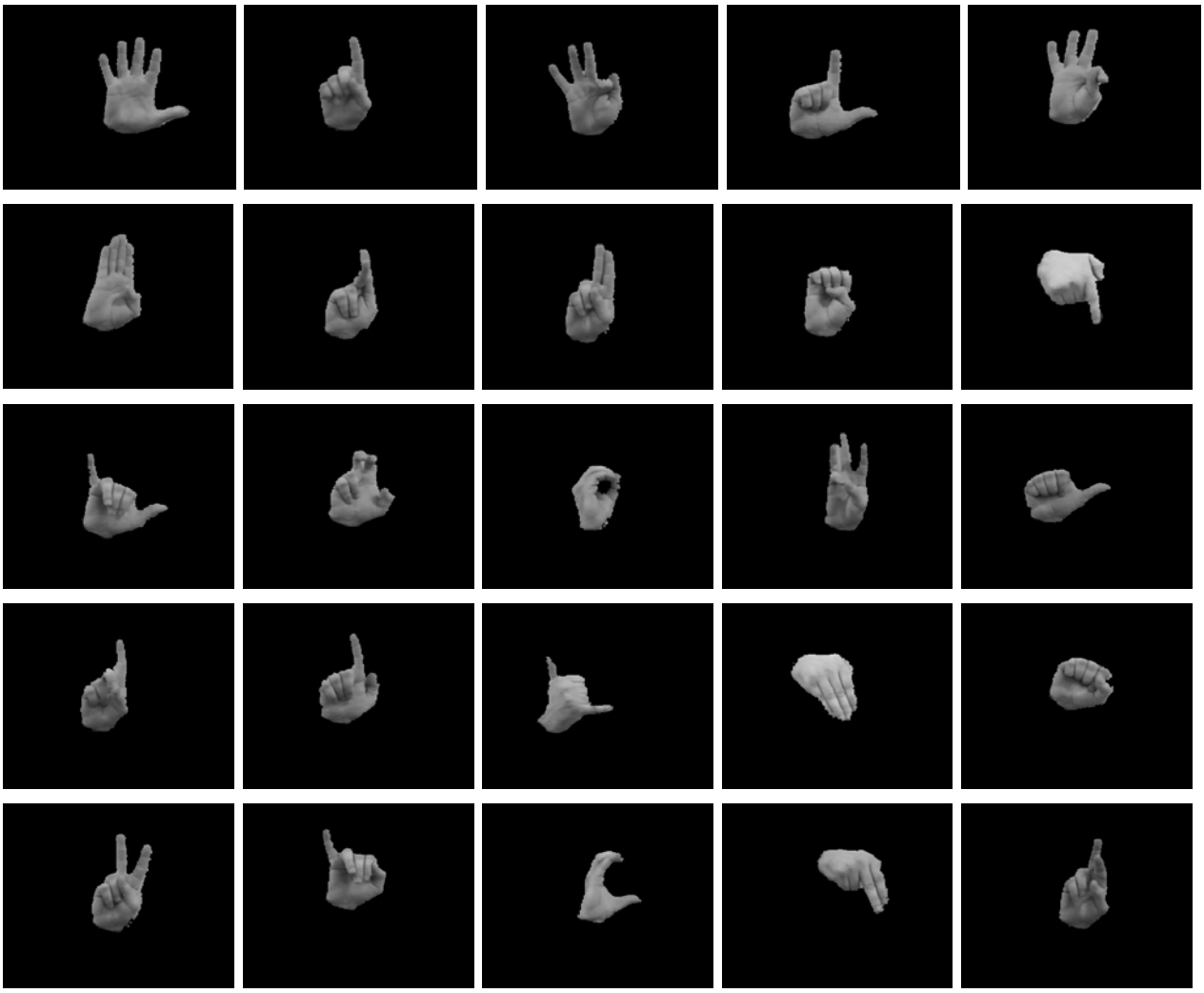


**Figure 4.19.** Les 25 prototypes naturels.

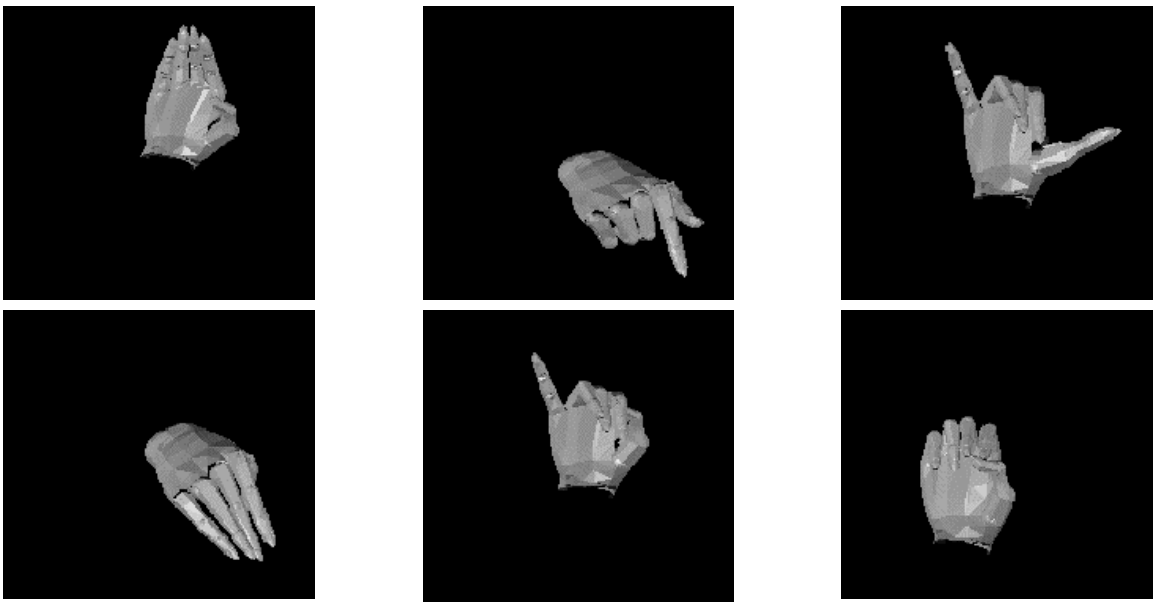


**Figure 4.20.** Quelques prototypes synthétiques.

Remarquons que ce concept de configuration stable doit toutefois intégrer une certaine variabilité par rapport à la configuration prototype. Cela est en effet indispensable, puisque tous les signeurs ne signent pas exactement de même façon et un même signeur ne reproduit jamais à l'identique un signe.

La correspondance entre identifiant de la séquence de geste, numéro de prototype et lettre associée est résumée dans le Tableau 4.3.

| **Identifiant de la séquence de geste** | **Numéro de prototype** | **Lettre** | **Identifiant de la séquence de geste** | **Numéro de prototype** | **Lettre** |
|---|---|---|---|---|---|
| Lettre 1 | 1 | Z | Lettre 15 | 14 | A |
| Lettre 2 | 2 | F | Lettre 16 | 15 | D |
| Lettre 3 | 3 | L | Lettre 17 | 16 | G |
| Lettre 4 | 4 | T | Lettre 18 | 17 | J |
| Lettre 5 | 5 | B | Lettre 19 | 18 | M |
| Lettre 6 | 6 | H | Lettre 20 | 7 | P |
| Lettre 7 | 7 | U | Lettre 21 | 19 | S |
| Lettre 8 | 8 | E | Lettre 22 | 20 | V |
| Lettre 9 | 9 | Q | Lettre 23 | 21 | Y |
| Lettre 10 | 10 | I | Lettre 24 | 22 | C |
| Lettre 11 | 11 | X | Lettre 25 | 6 | K |
| Lettre 12 | 12 | O | Lettre 26 | 23 | N |
| Lettre 13 | 13 | W | Lettre 27 | 24 | R |
| Lettre 14 | 10 | J | Lettre 28 | 1 | Z |

**Tableau 4.3.** Correspondance entre identifiant de la séquence de geste, numéro de prototype et lettre associée.

Notons que des lettres différentes peuvent avoir le même prototype, comme c'est le cas pour les lettres "P" et "U", et "K" et "H". Toutefois, pour ces dernières ("K" et "H"), les configurations associées sont en fait légèrement différentes, mais leurs projections dans l'image 2D sont pratiquement identiques.
Dans de tels cas, seule l'information de mouvement permet de discriminer entre ces lettres.

La deuxième étape de notre approche concerne la segmentation et le suivi de la main, que nous présentons par la suite.

#### 4.3.2.2 Segmentation et suivi de la main

La procédure de segmentation de la main adoptée, décrite dans [Mozelle99, Mozelle98T], combine apprentissage de teintes de chair et morphologie mathématique et consiste en les étapes suivantes :

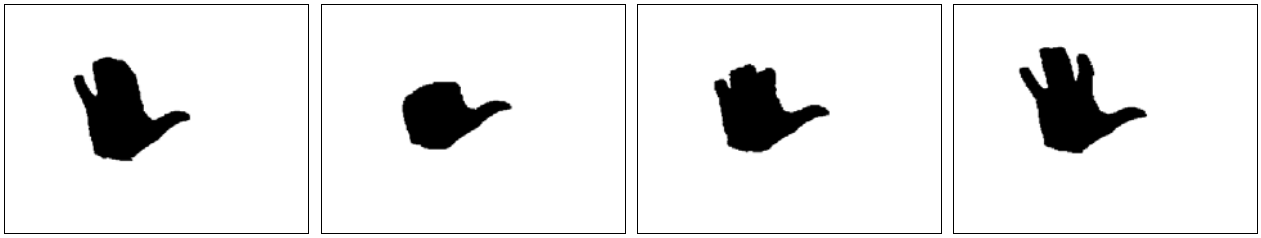
- apprentissage initial des teintes de chair de la main, effectué de manière interactive à l'aide de la plate-forme AMIS. L'utilisateur sélectionne plusieurs (quatre à cinq) *patchs* rectangulaires dans la séquence, correspondant à des régions de mains ou de visage dans différentes trames de la séquence. Un histogramme de couleur dans l'espace YUV est ensuite construit spécifiant le domaine des couleurs de teinte de chair. Un test d'appartenance à ce domaine fournit alors un premier masque de

segmentation, qui sera utilisé ensuite comme marquage.

- calcul du coût de connexion [Prêteux92] du gradient de l'image initiale [Canny86] par rapport aux marqueurs et extraction de la ligne des partage des eaux [Prêteux92].
- Etiquettage des différentes composantes détectées en main gauche, main droite et tête.

La Figure 4.21 montre quelques résultats de segmentation de la main droite.

a. Séquence initiale.



b. Masques de segmentation.

**Figure 4.21.** Segmentation de la main droite.

Sur les deux corpus utilisés dans notre travail, cette procédure produit des résultats de segmentation corrects avec un taux de succès de 99%.

Cette procédure de segmentation fournit donc pour chaque trame un masque binaire 2D de la main, que nous nous proposons d'analyser et de caractériser à l'aide de descripteurs de forme.

### 4.3.2.3 Descripteurs de configuration de la main

Dans ce paragraphe, nous considérons plusieurs descripteurs de configuration de la main : descripteurs de Hough 2D et de Fourier que nous avons initialement utilisés lors de nos recherches [Zaharia99], et descripteurs de forme actuellement retenus dans MPEG-7 (*cf.* Chapitre 1). Nous en discuterons en détails les différents aspects d'invariance aux transformations de similarité et les performances finales par rapport à l'application spécifique ici ciblée.

#### 4.3.2.3.1 La Transformée de Hough 2D

La transformée de Hough, détaillée au Chapitre 3 dans le contexte 3D, s'applique également dans un cadre 2D, en remplaçant l'ensemble des plans de la TH3D par un ensemble de droites dans le plan 2D, paramétrées en coordonnées polaires.

L'utilisation de la transformée de Hough 2D pour nos objectifs spécifiques de détection de configuration de main a été motivée par sa grande popularité dans le domaine de la détection des droites dans les images. En effet, les doigts de la main définissent en terme de région support des segments de droite plus ou moins longs, selon les différentes configurations. Un descripteur à base de transformée de Hough 2D (TH 2D) est donc *de facto* un bon candidat pour caractériser la configuration de la main.

Rappelons donc la définition et le principe de calcul de la TH 2D.

Soit $(xOy)$ un repère cartésien dans le plan 2D. Une droite L de $\mathbb{R}^2$, d'orientation $\theta$ et à distance $s$ de l'origine du repère (Figure 4.22) est caractérisée par :

$$L : \{ (x, y) \in \mathbb{R}^2 \mid s = x \cos\theta + y \sin\theta \}. \tag{4.1}$$

En faisant varier $s$ dans l'intervalle $[0,\infty)$ et $\theta$ dans $[0,2\pi)$, on obtient une paramétrisation en coordonnées polaires des droites de $\mathbb{R}^2$.

La TH 2D associe à chaque droite de $\mathbb{R}^2$ un point $(s,\theta)$ dans le domaine des coefficients de Hough (Figure 4.22).

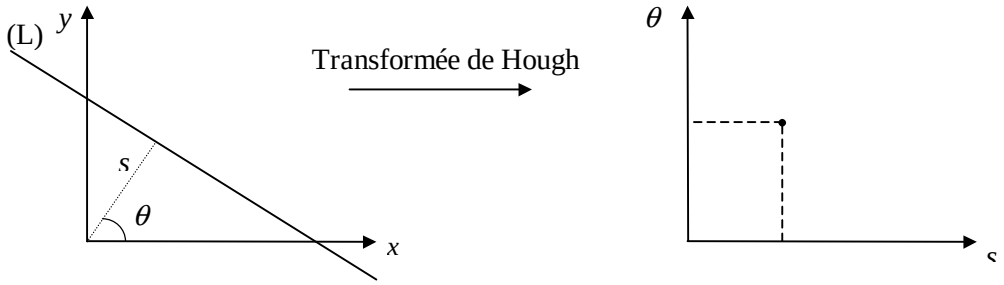



**Figure 4.22.** Le principe de la transformée de Hough 2D.

La TH 2D est liée à la transformée de Radon, dont on rappelle la formule intégrale ci-dessous (4.2) :

$$g(s,\theta)=\iint f(x,y)\delta(x\cos\theta+y\sin\theta-s)dxdy \tag{4.2}$$

où

- $f(x,y)$ est la fonction à analyser,
- $g(s,\theta)$ sa transformée de Radon, et
- $\delta$ désigne la distribution de Dirac.

La transformée de Radon intègre les valeurs de la fonction à analyser $f(x,y)$ sur l'ensemble des droites de $\mathbb{R}^2$, paramétré par $s$ et $\theta$. La réversibilité de la transformée de Radon, lorsque la totalité des projections est connue, assure un bon cadre pour la représentation du concept de forme qui perdure en pratique, même lorsqu'un nombre fini de projections est accessible.

Dans notre implantation, *f* correspond à la fonction support de la main, obtenue par l'algorithme de segmentation décrit précédemment. Les niveau de gris sont ici ignorés, afin d'éviter que les effets d'ombre, ou que les contours internes de la main n'influent sur les valeurs de la transformée.

En pratique, pour obtenir la TH 2D d'une fonction support binaire, nous appliquons une procédure d'accumulation similaire à celle définie pour la TH 3D au Chapitre 3.
Une partition uniforme du domaine des coefficients de Hough est tout d'abord effectuée en un nombre $N_s \times N_\theta$ d'intervalles de quantification, notés $Q_s$ et $Q_\theta$, et respectivement définis par la relation (4.3) :

$$Q_s = \{s_k = k\Delta s\}_{k \in \{0,1,\ldots,N_s-1\}}, \qquad Q_\theta = \{\theta_j = j\Delta\theta\}_{j \in \{0,1,\ldots,N_\theta-1\}} \tag{4.3}$$

où

- $\Delta\theta = \dfrac{2\pi}{N_\theta}$ désigne le pas de quantification de la coordonnée angulaire $\theta$ ,
- $\Delta s = \dfrac{S_{MAX}}{N_s}$ est le pas de quantification de la coordonnée *s*,
- $S_{MAX}$ définit l'échelle de l'objet et est considéré ici comme la dimension maximale de la main dans les images.

La procédure de calcul de la transformée de Hough, notée $\{h(k,j)\}_{\substack{k \in \{0,1,\ldots,N_s-1\} \\ j \in \{0,1,\ldots,N_\theta-1\}}}$, consiste en les étapes suivantes, décrites ci-dessous en pseudo-code :

*Initialisation* de $h(k, j)$ à 0, quels que soient *k* et *j* ;
**Pour** chaque pixel (*x*, *y*) du masque de l'objet ,
  **Pour** chaque orientation $\theta_j$ de $Q_\theta$ ,
    *Calcul* de *s* : $s = x\cos\theta_j + y\sin\theta_j$ ;
    **Si** *s* positif,
      *Quantification* de *s* au plus proche $s_k$ de $Q_s$ ;
      *Actualisation* de *h* : $h(k, j) = h(k, j) + 1$ ;
    **Fin Si**
  **Fin Pour** chaque orientation
**Fin Pour** chaque pixel

Le test sur le signe de *s* a ici les mêmes raisons que celles exposées au Chapitre 3, dans le cadre de la TH 3D.

Ainsi définie, la TH 2D n'est pas invariante aux transformations de similarité. Or, le même geste peut être signé dans différentes positions et la même configuration peut subir des transformations au cours du même geste.

Il est donc important d'examiner comment cette invariance peut être obtenue.

#### 4.3.2.3.2 La transformée de Hough 2D invariante aux similarités

L'invariance de la transformée de Hough aux translations est réalisée en positionnant l'origine du repère cartésien considéré au centre de gravité de la main.

Quant aux aspects d'invariance par rapport aux rotations, ils sont traités à l'aide d'une analyse en composantes principales :

- Les deux axes principaux d'inertie sont tout d'abord déterminés,
- L'axe des *x* est ensuite défini comme l'axe principal de plus grande valeur propre.

Comme l'objectif ici est la détection des gestes individuels et comme l'objet considéré est toujours la main, les problèmes complexes d'alignement traités au Chapitre 3 n'interviennent plus.
En revanche, l'invariance aux homothéties reste de la plus haute importance, puisqu'en pratique les signeurs peuvent être situés plus ou moins loin de la caméra. Les tailles de la main peuvent être alors très différentes. Pour ne mentionner que les deux corpus sur lesquels nous avons travaillé, "Lettres" et "Mots", la taille de la main varie d'un facteur 2 à 3 d'un corpus à l'autre.
Pour assurer un comportement invariant aux homothéties, rappelons une propriété bien connue de la transformée de Radon.
Si $f_a(x,y) = f(ax,ay)$ désigne une version homothétique de la fonction support $f(x,y)$ ($a$ étant un facteur d'homothétie positif), sa transformée de Radon, notée $g_{(a)}(s,\theta)$ s'exprime en fonction de celle de la fonction initiale, $g(s,\theta)$, par :

$$g_{(a)}(s,\theta) = \frac{1}{a} g(as,\theta) \tag{4.4}$$

La fonction $g_{(a)}(s,\theta)$ s'exprime donc à partir de $g(s,\theta)$ par une homothétie d'un même facteur *a* selon la variable *s* et une mise à l'échelle supplémentaire, d'un facteur $\frac{1}{a}$, de ses valeurs.
Cette propriété, illustrée Figure 4.23, permet de dériver une version invariante à l'échelle de la transformée de Hough, notée $\overset{\circ}{h}(k,j)$, et définie à partir de $h(k,j)$ par :

$$\overset{\circ}{h}(k,j) \overset{def}{=} \frac{1}{\lambda} h(\lambda k, j), \tag{4.5}$$

où

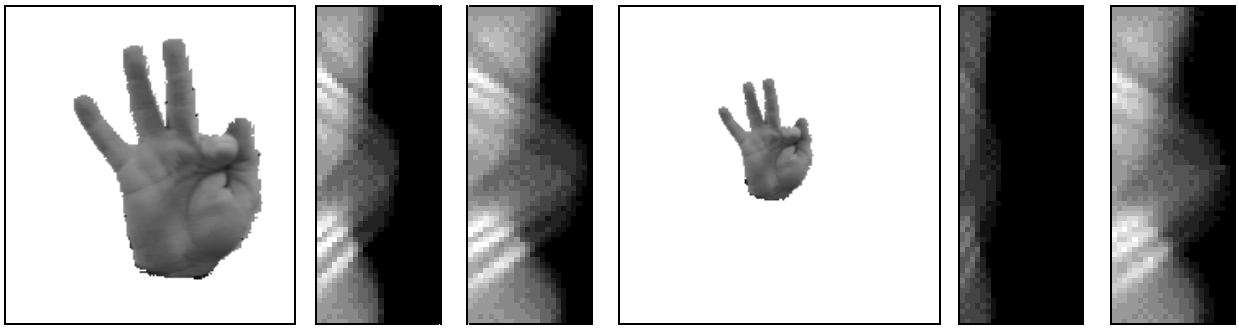
$\lambda = \frac{N_s}{k_{\max}}$, avec $k_{\max} = \max\{k \mid h(k,j) \neq 0\}$.

Notons que $\lambda k$ est un nombre réel, en général non-entier. Pour cette raison, afin de déterminer les valeurs de $\overset{\circ}{h}(k,j)$, nous effectuons une interpolation bilinéaire à partir des valeurs entières les plus proches de $\lambda k$, comme suit :

$$\overset{\circ}{h}(k,\,j) = (1-\alpha_{\lambda k})h([\lambda k],\,j) + \alpha_{\lambda k}\,h([\lambda k]+1,\,j)\,, \tag{4.6}$$

où $[\cdot]$ désigne la partie entière et $\alpha_{\lambda k}$ la partie fractionnaire de $\lambda k$.

Pour simuler la façon dont l'invariance par rapport à l'échelle est obtenue, nous avons généré à partir d'une configuration donnée, sa version homothétique pour un facteur d'échelle 1 : 2. La Figure 4.23 montre les deux images de configuration ainsi produites, les transformées de Hough 2D associées et la transformée de Hough invariante.



a. Configuration initiale, transformée de Hough et sa version invariante.

a. Homothétie de facteur 1 : 2 de la configuration a, transformée de Hough et sa version invariante.

**Figure 4.23.** Illustration du principe de la Transformée de Hough invariante à l'échelle.

La Figure 4.24 montre les transformées de Hough associées aux 25 prototypes de la Figure 4.19. Les pas de quantification sont ici $\Delta\theta = 0.1$ radians et $\Delta s = 5$ pixels.

Pour interpréter ces images de Hough, rappelons qu'un point arbitraire (*x*, *y*) du domaine donne des contributions sur une sinusoïde dans le plan $(s,\theta)$, décrite par l'équation (4.7) :

$$s = r\cos(\theta-\varphi)\,, \tag{4.7}$$

où *r* et $\varphi$ désignent les coordonnées polaires du point (*x*, *y*).

La transformée de Hough d'un objet est en effet obtenue par la superposition de telles fonctions sinusoïdales. Les différentes accumulations visibles Figure 4.24 proviennent de l'intersection des sinusoïdes ayant la même phase $\varphi$ et des amplitudes *r* variables. Elles correspondent en effet aux points du masque distribués spatialement selon des segments de droite qui sont, dans notre cas, les doigts. Ces accumulations sont bien localisées dans le domaine $(s,\theta)$ lorsque les doigts sont bien écartés et prennent une extension spatiale plus importante dans le cas de configurations avec un ou plusieurs doigts collés, en raison de la fusion de deux maxima adjacents en un seul.

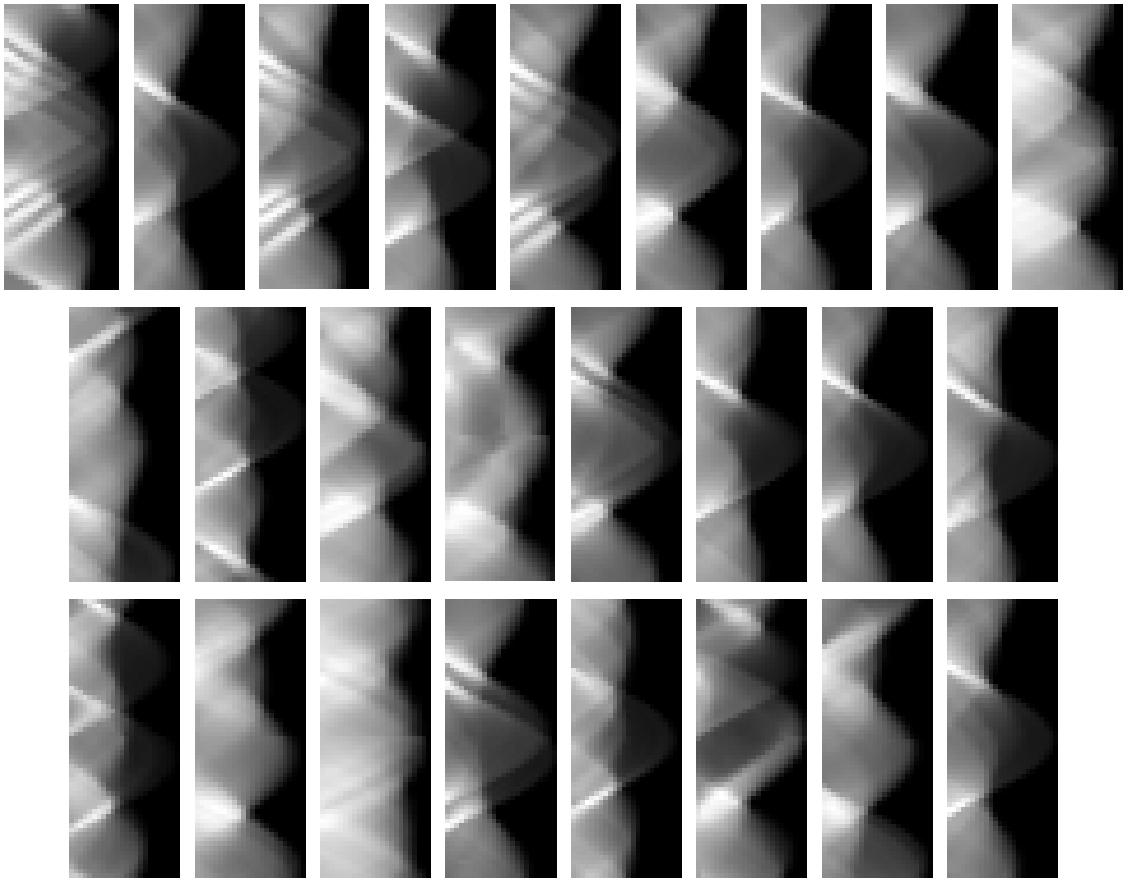


**Figure 4.24.** Transformées de Hough invariantes associées aux 25 prototypes.

La transformée de Hough, dans sa version invariante, offre donc un outil adapté pour représenter l'anisotropie des différentes configurations de main.
La Figure 4.24 montre que les transformées de Hough présentent néanmoins un haut degré de corrélation. Etudions donc comment décorréler cette représentation, afin d'obtenir une représentation plus compacte.

#### 4.3.2.3.3 Descripteur de configuration à base de transformée de Fourier

Ce descripteur de configuration est dérivé de la TH 2D en lui appliquant la transformée de Fourier.
Comme la transformée de Hough représente des signaux 2D à valeurs réelles, sa transformée de Fourier sera alors conjuguée et symétrique par rapport à l'origine du domaine des fréquences spatiales. Il en résulte que seule la moitié des coefficients de Fourier, correspondant à un demi-plan du domaine spectral, suffit pour obtenir une description complète de la transformée de Hough. Comme ces coefficients sont des nombres complexes, la complexité du descripteur reste globalement la même.

Pour donner une interprétation mathématique du descripteur de Fourier, rappelons le théorème de projection [Gindikin94]. Il établit que la représentation polaire de la fonction initiale, notée $F_p(\xi,\theta)$ , et donnée par :

$$F_p(\xi,\theta)=F(\xi\cos\theta,\xi\sin\theta)\,, \tag{4.8}$$

est en fait la transformée de Fourier 1D (intégration selon la variable *s*) de la transformée de Radon $g(s,\theta)$ en $s = \xi$ .

Il en résulte, en utilisant la séparabilité de la transformée de Fourier, que le descripteur par transformée de Fourier, noté $G(\omega_s,\omega_\theta)$ , est donné par :

$$G(\omega_s,\omega_\theta)=Fourier_{2D}\{g(s,\theta)\}(\omega_s,\omega_\theta)= \underset{\theta}{Fourier_{1D}}\left\{F_p(\xi,\theta)\big|_{\xi=\omega_s}\right\}(\omega_\theta)\,, \tag{4.9}$$

ce qui correspond à une construction similaire à celle de la transformée de Fourier-Mellin modulo la représentation logarithmique de *s*.

Les valeurs absolues des coefficients de Fourier correspondant aux 25 configurations de main considérées sont illustrées Figure 4.25.

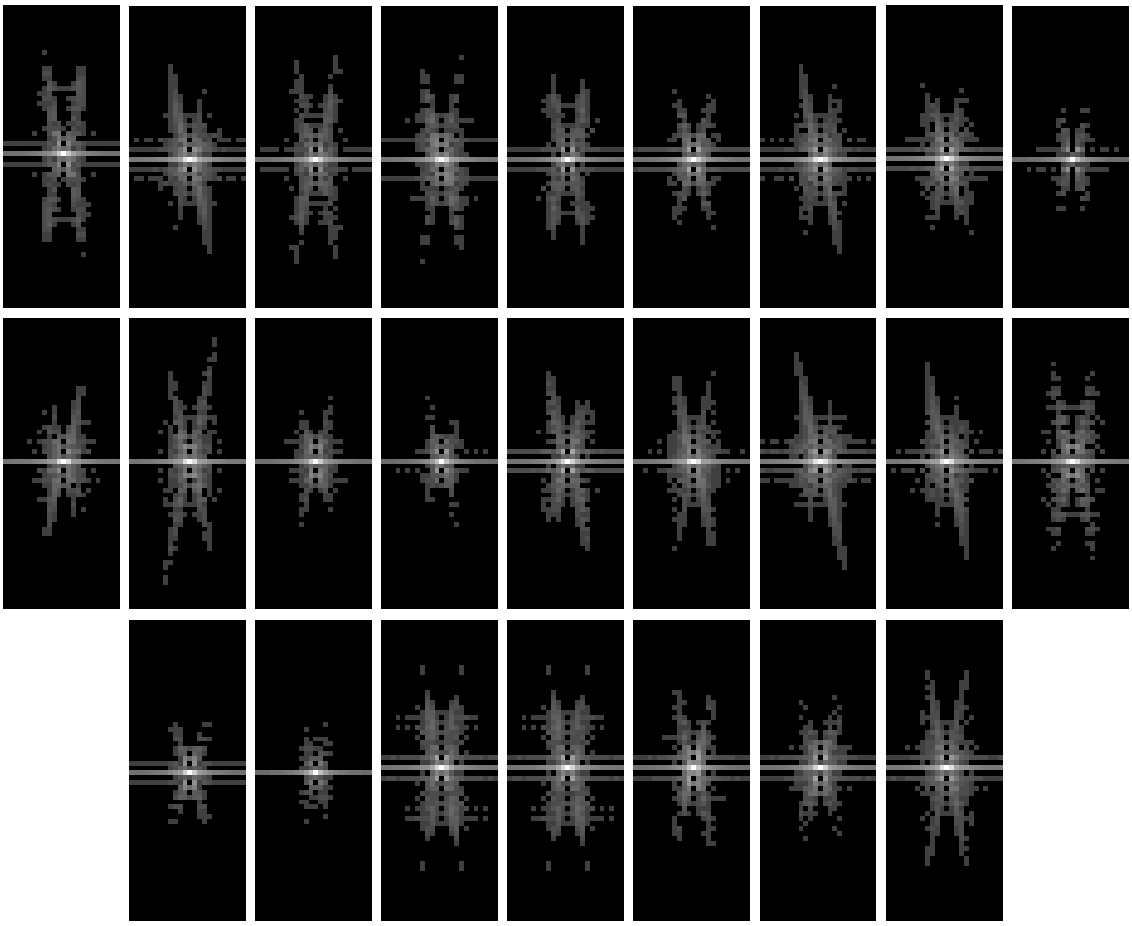


**Figure 4.25.** Valeurs absolues des coefficients de Fourier correspondant aux 25 prototypes de configuration.

La Figure 4.25 montre qu'un nombre réduit de coefficients de Fourier, correspondant aux basses fréquences, a des valeurs significatives. En pratique, nous retenons le quart des coefficients, en seuillant à 0

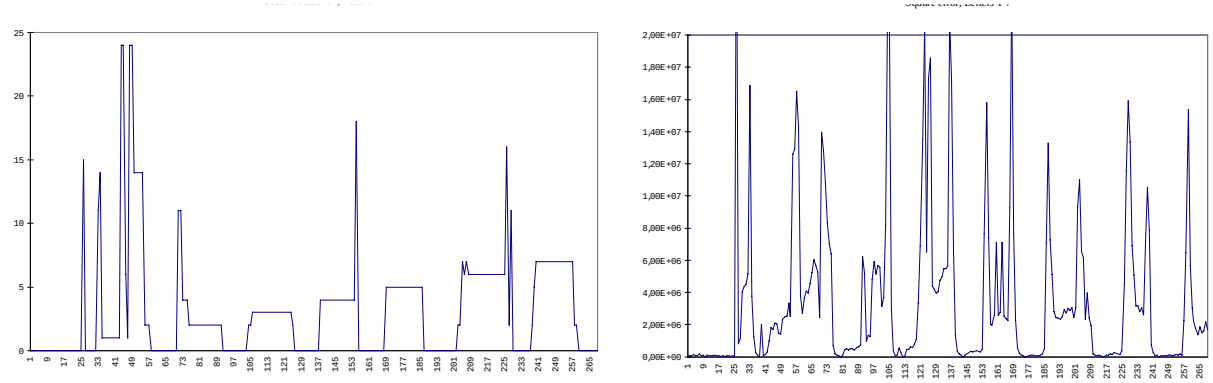
les composantes du spectre de fréquences supérieures à la moitié de celles maximales selon au moins l'un des deux axes. La complexité de la représentation est donc le quart de celle initiale.

En outre, remarquons que les valeurs absolues des coefficients de Fourier sont invariantes aux rotations dans le domaine initial (*x*, *y*) de la forme, puisque celles-ci se traduisent en terme de transformée de Hough par un changement de phase $\theta$.
En prenant une distance $L_1$ entre les valeurs absolues des coefficients de Fourier, nous obtenons donc une mesure de similarité intrinsèquement invariante aux rotations.

#### 4.3.2.4 Résultats expérimentaux

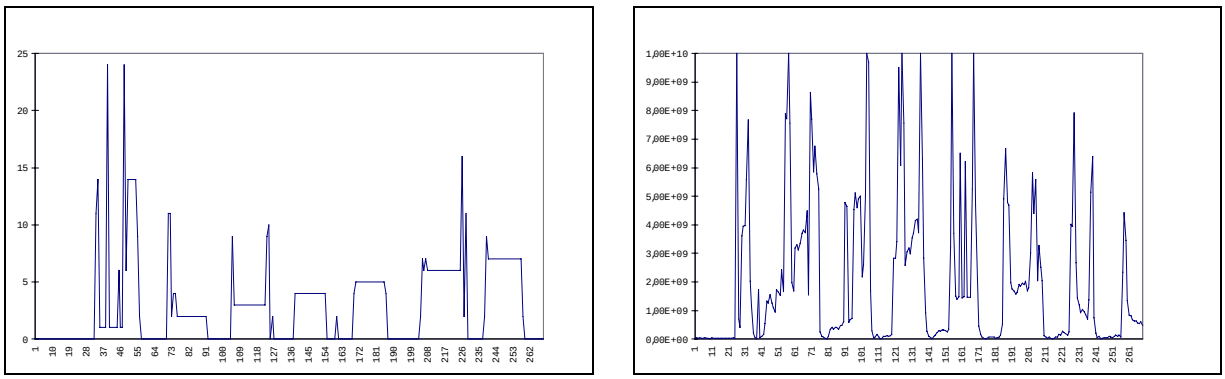
Afin de tester la pertinence des descripteurs de configuration proposés, nous avons conduit trois expérimentations différentes.
Dans la première, nous avons considéré uniquement le corpus "Lettres". Pour chaque trame des séquences vidéos, la configuration de la main est comparée à chaque configuration de l'ensemble des prototypes, en utilisant une distance $L_1$ entre les descripteurs de Hough (et de Fourier). Le prototype donnant la distance minimale est sélectionné comme configuration détectée.

Les Figures (Figure 4.26 et Figure 4.27) montrent quelques résultats de détection et les distances associées au prototype le plus proche.



a. Configurations détectées. b. Distances au plus proche prototype.

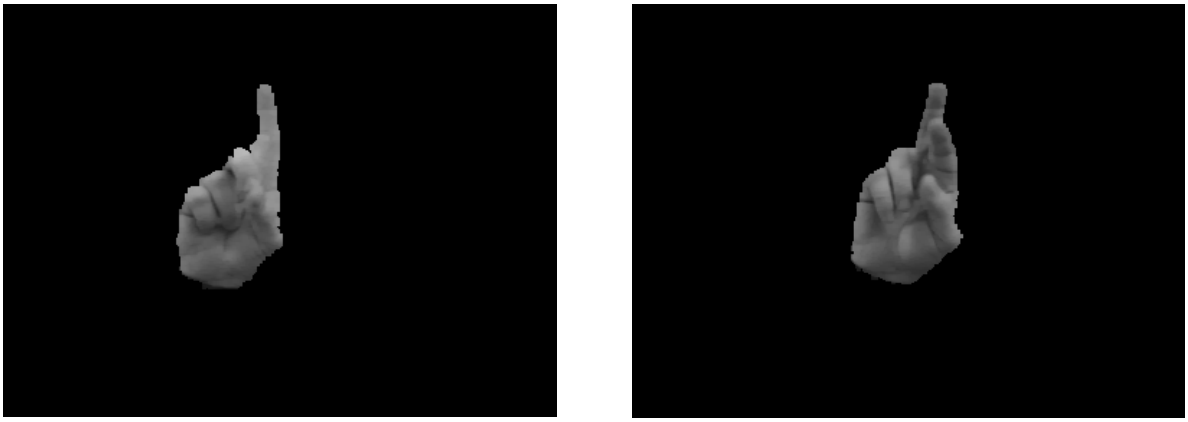
**Figure 4.26.** Reconnaissance de gestes par descripteur de Hough 2D.



a. Configurations détectées. b. Distances au plus proche prototype.

**Figure 4.27.** Reconnaissance de gestes par descripteur de Fourier.

Remarquons que deux configurations différentes, correspondant à un "Z" et à un "R" sont détectées de manière prépondérante dans la séquence "Lettre 1" (signe "Z"). Cela s'explique par la grande similarité de forme 2D entre celles-ci (Figure 4.28).

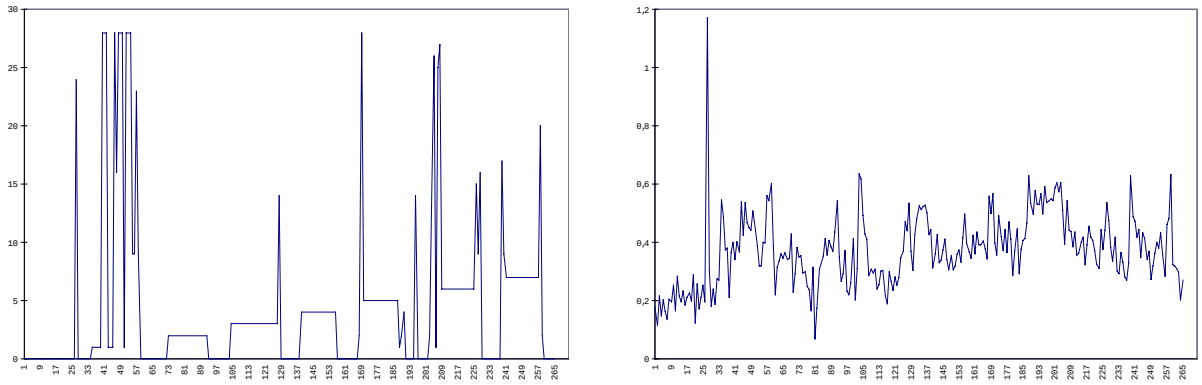


a. Lettre "Z". a. Lettre "R".

**Figure 4.28.** Configurations des lettres "Z" et "R": les projections 2D de la main sont presque identiques.

Afin d'évaluer objectivement les résultats de reconnaissance, nous avons sélectionné manuellement et retenu pour chaque lettre, les trames correspondant à des configurations stable, en éliminant les trames de transition de configuration. Le taux de reconnaissance obtenu est de 98%.

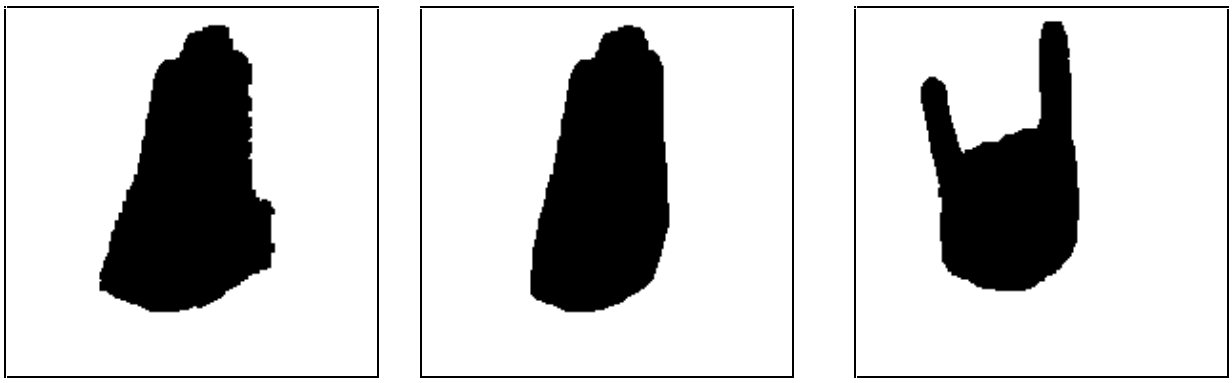
Afin d'étudier l'impact sur les performances de détéction des descripteurs de forme MPEG-7, nous avons répliqué l'expérimentation, en utilisant cette fois le descripteur MPEG-7 d'espace-échelle de contour (Chapitre 1). Ce choix est justifié par l'observation que les configurations peuvent être représentées en général (à l'exception de la lettre *O*) par un seul contour fermé, les évaluations conduites au sein de MPEG-7 ayant démontré dans ce cas, la nette supériorité de ce descripteur par rapport aux descripteurs MPEG-7 à base de régions.
La Figure 4.29 montre, pour les séquences de gestes de 1 à 7, que les bons résultats en terme de détection se maintiennent. Toutefois, les courbes des distances au plus proche prototype montrent un comportement beaucoup moins fiable en terme de pouvoir de discrimination.



a. Configurations détectées. b. Distances au plus proche prototype.

**Figure 4.29.** Reconnaissance de gestes par le descripteur MPEG-7 d'espace échelle de contour.

La deuxième expérimentation, a pour objectif de tester l'invariance à l'échelle des signatures de forme proposées, les prototypes naturels sélectionnés à partir des séquences "Lettres", étant ici utlilisés pour la reconnaissance des configurations acquises en cadrage large caractéristique du corpus "Mots". Notons que trois prototypes complémentaires ont dû être ajoutés ici (Figure 4.30), un pour le signe "Tiret", et deux autres pour des versions non-standard des signes "B" et "H", n'apparaissant pas dans le corpus "Lettres", mais utilisées dans le corpus "Mots".



a. Configuration "Tiret" b. Configuration non-standard de "B". c. Configuration non-standard de "H".

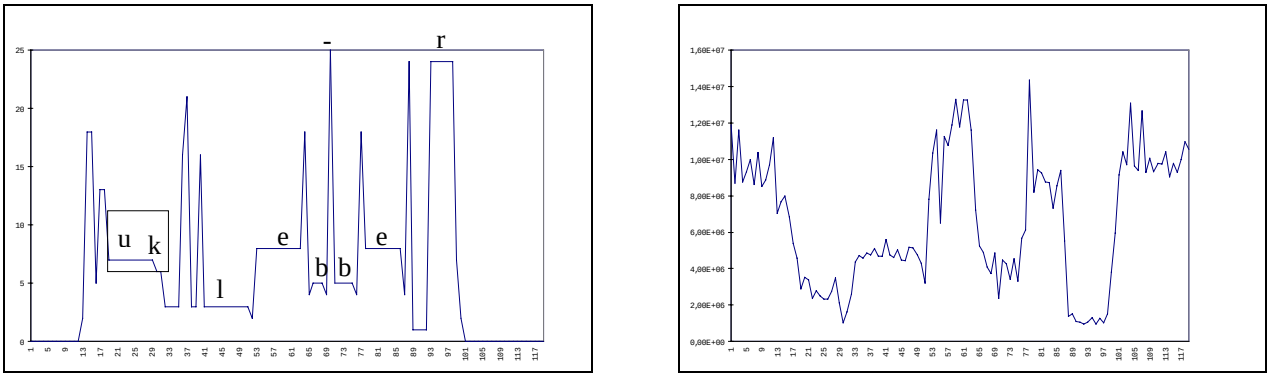
**Figure 4.30.** Prototypes supplémentaires correspondant aux configurations nouvelles apparaissant dans le corpus "Mots".

Les résultats de reconnaissance sont présentés Figures (Figure 4.31 et Figure 4.32). En général, la configuration correcte est détectée, malgré la forte variation du facteur d'échelle entre les prototypes utilisés et les configurations du corpus "Mots".

Toutefois, quelques fausses détections apparaissent ici.

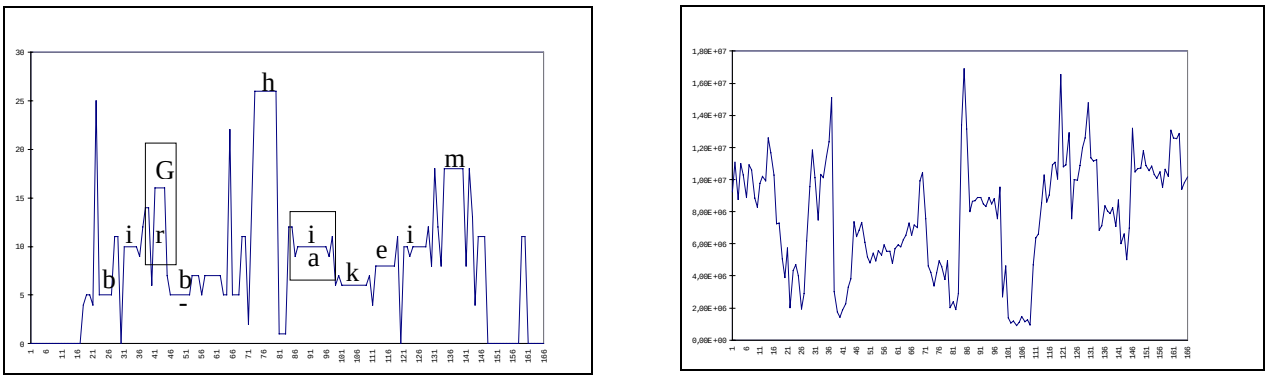
Ainsi, dans la séquence "Kleber", la lettre "K" est reconnue comme un "U". Cela s'explique par le fait que, même si les positions 3D de main correspondant à ces configurations sont différentes, leurs projections dans l'image 2D sont presque identiques (d'autant plus que dans le cas de ce corpus les tailles de main sont très petites, les éventuelles différences devenant négligeables).

Le même problème, dû aux mêmes raisons, apparaît avec les signes "B" et "-".




a. Configurations détectées. b. Distances au plus proche prototype.

**Figure 4.31.** Résultats de la reconnaissance de gestes (descripteur de Hough) sur la séquence "Kleber".




a. Configurations détectées. b. Distances au plus proche prototype.

**Figure 4.32.** Reconnaissance de gestes (descripteur de Hough) sur la séquence "Bir-Hakeim".

Dans la séquence "Bir-Hakeim", la lettre "R" est reconnue comme "G". En outre, observons que la lettre "A" est reconnue comme "I". Cela est dû au fait que le geste "A" rencontré ici est signé de manière non-standard, le prototype associé ne pouvant donc pas être reconnu, puisqu'il n'est pas inclus dans la base des prototypes.

Ces deux premières expérimentations prouvent la capacité des deux descripteurs considérés à discriminer entre les différentes configurations de la main. Toutefois, elles ont été conduites en n'utilisant que les prototypes naturels et jamais ceux de synthèse. Qu'advient-il de ces performances quand l'indexation exploite ces derniers ?
Nous avons repliqué la première expérimentation sur le corpus "Lettres" mais en utilisant les prototypes de synthèse. Le taux de reconnaissance exacte a alors drastiquement chuté à 30%.

Cette contre-performance est liée au fait que la main synthétique utilisée a des caractéristiques morphométriques (longueur et taille relatives des doigts et de la paume) très différentes de celles correspondant aux mains réelles dans les séquences vidéos. Afin de pallier cet inconvénient, une calibration morphométrique de la main synthétique a été réalisée (Figure 4.33), de telle sorte que celle-ci reflète mieux

les caractéristiques de la main réelle. Le taux de reconnaissance correcte est alors remonté d'une manière spectaculaire à 89%. Cela démontre bien que la morphométrie est un attribut essentiel qui doit être pris en compte de manière appropriée dans un contexte de reconnaissance.
En outre, remarquons que cette reconnaissance à base de prototype synthétique offre l'avantage de la connaissance des paramètres 3D de la main.

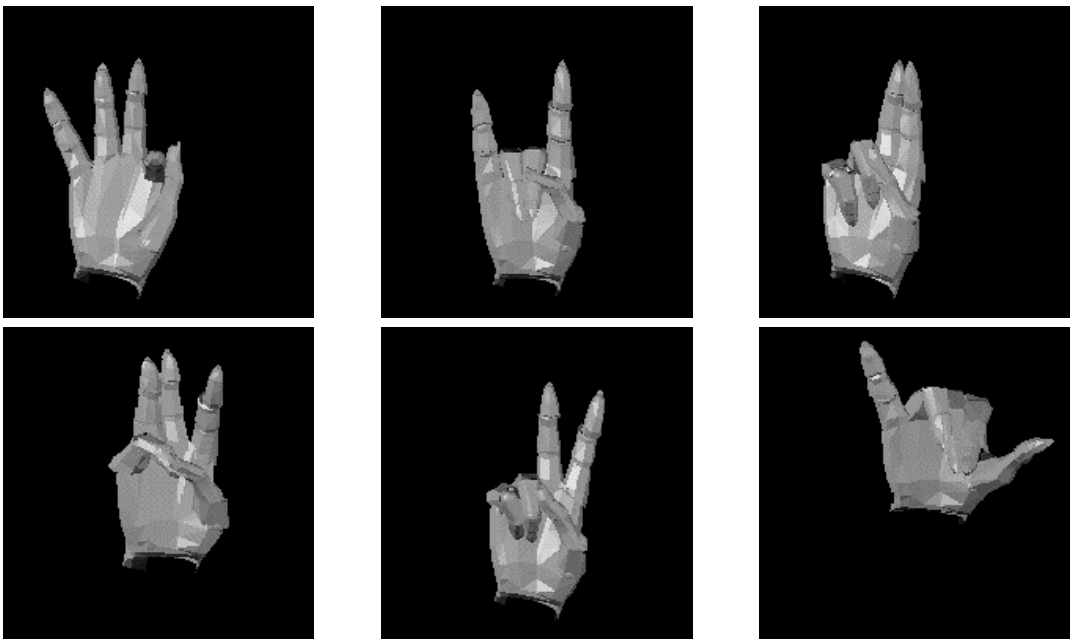


**Figure 4.33.** Prototypes synthétiques calibrésen fonction de la morphométrie du signeur.
(à comparer avec les prototypes non-calibrés de la Figure 4.20).

Les similitudes de forme existant entre différentes configurations de main montrent que, pour obtenir une meilleure reconnaissance, il est toutefois nécessaire de considérer les autres caractéristiques énoncées par les experts en langue des signes. Etudions donc à présent comment ces différents attributs de geste peuvent être intégrés dans un schéma de description unifié et comment les structures actuellement retenues au niveau de MPEG-7 peuvent supporter une telle description.

#### 4.3.2.5 Reconnaissance de la langue des signes dans le contexte MPEG-7

Le schéma de description région spatio-temporelle (*MovingRegion DS* – MR DS) nous paraît suffisamment générique et particulièrement bien adapté pour prendre en compte l'hétérogénéité des différentes caractéristiques des gestes.
Comme nous l'avons vu au Paragraphe 4.2, le MR DS fédère des descripteurs d'image fixe, de mouvement, de localisation et d'annotation textuelle. De plus, le mécanisme de décomposition spatio-temporelle hiérarchique permet de gérer de façon naturelle les différents objets d'intérêt (typiquement la tête et les deux mains), et leur décomposition en sous-segments, comme par exemple en segment de configuration stable ou de mouvement affine stable.
Les relations spatiales entre ces différents objets sont quant à elles prises en compte par le *SegmentRelation DS*.

Le nombre de mains utilisées pour signer, l'orientation des doigts et de la paume peuvent être intégrés comme simples annotations textuelles, comme elles peuvent être plutôt utilisées pour des applications de filtrage.
Les modèles de mouvement paramétrique et les mécanismes de segmentation temporelle par le mouvement présentés au Chapitre 2 permettent de décrire efficacement la notion d'action.
Quant à la notion de configuration, nous avons déjà montré que le descripteur d'espace-échelle de contour fournit des performances raisonnables en terme de détection.

La Figure 4.34 résume les éléments de MPEG-7 que nous avons retenus et montre la façon dont ils acceuillent les différents descripteurs spécifiques de la langue des signes.

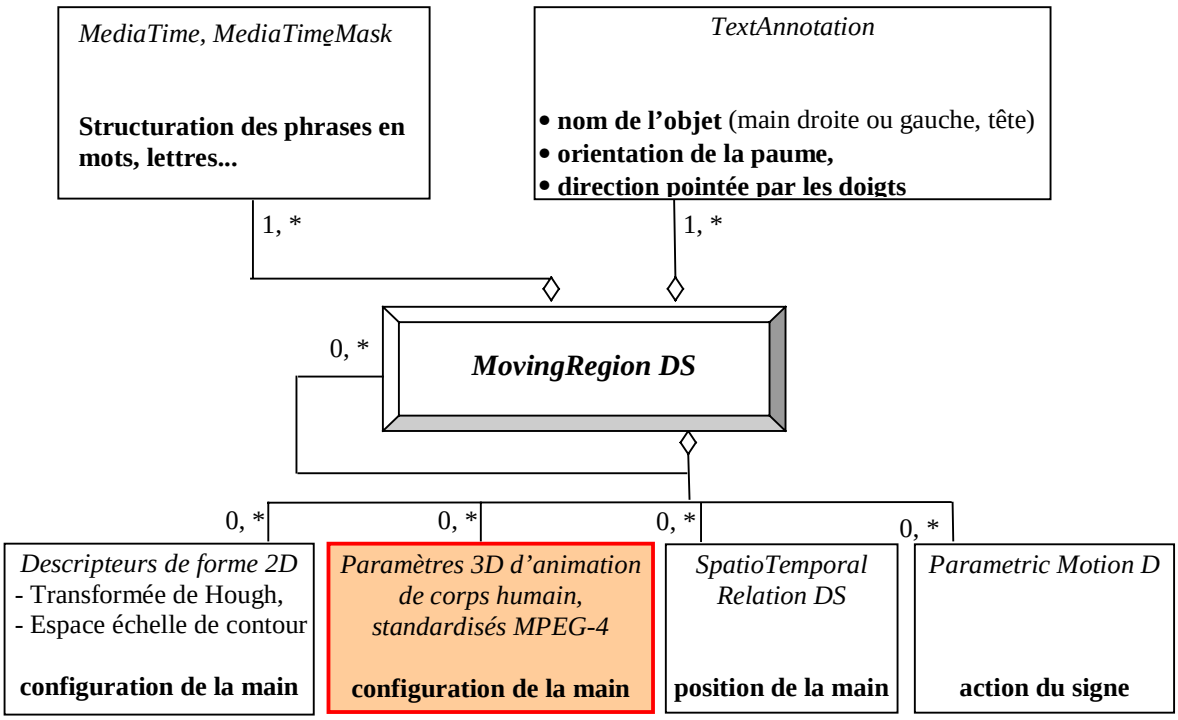



**Figure 4.34.** Eléments MPEG-4 et MPEG-7 pour l'indexation de la langue des signes.

Les différents connecteurs de la Figure 4.34 sont représentés en notation UML (Figure 4.35) [UML].

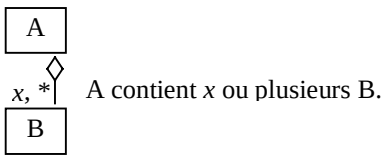



**Figure 4.35.** Notation UML.

Nous avons ajouté ici un module extérieur au *MovingRegion DS* qui correspond aux paramètres MPEG-4 d'animation du corps humain, puisque ceux-ci décrivent d'une manière complète la configuration de la main dans l'espace 3D. Rappelons qu'il existe actuellement dans MPEG-7 un schéma de description 3D,

appelé *StillRegion3D DS* qui permet d'intégrer des descripteurs de forme 3D. Définir de manière analogue un schéma de description *MovingRegion3D DS* permettrait d'élaborer des applications d'indexation de la langue des signes dans le contexte MPEG-7, tout en assurant une pleine compatibilité entre les standards MPEG-7 et MPEG-4 au niveau des médias supportés.

## 4.4 Conclusion

Ce chapitre propose une plate-forme d'indexation compatible MPEG-7 dédiée à la visualisation/navigation/annotation de contenus audio-visuels naturels et synthétiques, exploitant notamment les avancées méthodologiques en termes de schémas de description réalisées au sein de MPEG-7. Fondé sur une architecture modulaire extensible, le système intègre divers types de médias (images fixes, séquences d'images, documents vidéos, maillages 3D) sous différents formats et offre un ensemble diversifié d'outils de segmentation temporelle et spatio-temporelle et d'extraction des descripteurs MPEG-7. En particulier, le principe d'interopérabilité devient effectif par la mise en œuvre des schémas de description MPEG-7 qui permettent de s'affranchir des problèmes de cohérence, d'héritage, de hiérarchie, de synchronisation et d'intégration de signatures hétérogènes.

L'ensemble logiciel ainsi réalisé démontre pour la première fois en grandeur réelle, pour les applications d'indexation multimédia, comme l'archivage vidéo, la vidéo cliquable et la reconnaissance des gestes en langue des signes française, le caractère effectivement opérationnel de schémas de description audio-visuels génériques, normalisés MPEG-7.

# 5 Conclusion et perspectives

Né au début de la dernière décennie du vingtième siècle avec l'avènement de l'Internet et l'accroissement du volume de données numériques, le domaine de l'indexation par le contenu des documents audio-visuels se voit aujourd'hui reconnu par la première version du standard MPEG-7 de description des données multimédias. Cette norme regroupe un ensemble de technologies ayant atteint un certain degré de maturité. L'analyse de celles-ci, effectuée au premier chapitre de ce mémoire, démontre que l'état de l'art dans le domaine est bien représenté dans MPEG-7, qui offre un solide support pour de nombreuses applications potentielles.

S'inscrivant pleinement dans le cadre du processus de standardisation MPEG-7, auquel nous avons participé activement sous forme de 24 propositions ISO (*cf.* Liste des publications associées), cette thèse aborde des aspects d'indexation encore peu discutés et propose :

- de nouvelles mesures de similarité pour les descriptions de mouvement par modèles paramétriques,
- deux descripteurs de forme pour indexer des modèles 3D maillés,
- une plate-forme d'indexation compatible MPEG-7, intégrant des outils d'annotation, de navigation, de visualisation et de requêtes par similarité, et supportant des applications comme l'archivage vidéo, la vidéo cliquable ou l'indexation MPEG-7 de la langue des signes française.

Malgré la compacité et la fidélité des représentations de mouvement par modélisation paramétrique, leur exploitation directe dans le cadre des applications de requêtes par similarité est restée peu explorée, par manque de mesures de similarité appropriées et de mécanismes de gestion du caractère dynamiques des données vidéos. Afin de lever ce verrou technologique, nous avons proposé une famille de mesures de similarité, définie dans l'espace des champs de vitesse, dépendant d'une fonction distance générique à spécifier en fonction du type de requête. Les problèmes d'alignement spatio-temporel et de pondération des composantes translationnelle et homogène de mouvement sont analysés et une solution mathématique proposée et mise en œuvre sous forme d'une famille de mesures de similarité simplifiées, optimisées en terme de complexité de calcul.
Les résultats obtenus démontrent objectivement et quantitativement par estimation du critère *Bull Eye*, la nette supériorité (avec un gain allant jusqu'à 25%) des mesures proposées par rapport aux mesures de similarité dans l'espace des paramètres, habituellement utilisées jusqu'à présent, et établit ainsi leur pertinence.
Cet ensemble de développements apporte une solution méthodologique tant théorique que pratique, son degré de maturité ne nécessitant plus de recherche nouvelle.

Ensuite, nous avons abordé le domaine encore très jeune de l'indexation par le contenu des maillages 3D, omniprésents aujourd'hui dans le monde de la graphique et de la réalité virtuelle. L'analyse de ces données complexes doit surmonter un certain nombre de difficultés d'ordre géométrico-topologique et sémantique, que nous avons discutées en détails.

En particulier, nous avons présenté, analysé et comparé deux approches différentes pour représenter des modèles polygonaux 3D, visant des applications de requête par similarité de forme.
Partant des propriétés d'invariance géométrique et topologique que doit satisfaire naturellement un DF d'objet 3D maillé, nous avons proposé tout d'abord le SF3D. Celui-ci exploite uniquement la structure géométrique locale d'une surface 3D, fournit une représentation très compacte mais offre des performances moyennes en terme de requête par similarité en raison d'une part de la perte de toute information de localisation spatiale et d'autre part de sa grande sensibilité aux descriptions topologiques des maillages.
Ensuite, en considérant la transformée de Hough 3D d'un maillage, nous avons construit le DH3DO, intrinséquement invariant aux changements de connexité, rendu indépendant des transformations géométriques et optimisé en terme de compacité de représentation, via une partition invariante aux changements de repère canonique de la sphère unité. Ses performances, excellentes en terme de score *Bull-Eye* (81%) sur une base catégorisée de 362 modèles, se conservent sur l'ensemble des 1300 modèles de la base MPEG-7, confirmant les remarquables propriétés de scalabilité du DH3DO.
Les perspectives méthodologiques de ce travail concernent l'élaboration d'un descripteur de forme hybride, combinant descriptions locales et globales des maillages, intrinsèquement invariant géométriquement et topologiquement afin de s'affranchir de toute procédure d'alignement spatial.

Enfin, le dernier chapitre propose une plate-forme d'indexation compatible MPEG-7 dédiée à la visualisation/navigation/annotation de contenus audio-visuels naturels et synthétiques, exploitant notamment les avancées méthodologiques en termes de schémas de description réalisées au sein de MPEG-7. Fondé sur une architecture modulaire extensible, le système intègre des divers types de médias (images fixes, séquences d'images, documents vidéos, maillages 3D) sous différents formats et offre un ensemble diversifié d'outils de segmentation temporelle et spatio-temporelle et d'extraction des descripteurs MPEG-7. En particulier, le principe d'interopérabilité devient effectif par la mise en œuvre des schémas de description MPEG-7 qui permettent de s'affranchir des problèmes de cohérence, d'héritage, de hiérarchie, de synchronisation et d'intégration de signatures hétérogènes.

L'ensemble logiciel ainsi réalisé démontre pour la première fois en grandeur réelle, pour les applications d'indexation multimédia, comme l'archivage vidéo, la vidéo cliquable et la reconnaissance des gestes en langue des signes française le caractère effectivement opérationnel de schémas de description audio-visuels génériques, normalisés MPEG-7.

Les perspectives de travail concernent principalement l'application de la reconnaissance des gestes, notamment par le développement, en coopération avec les experts en langue des signes, de dictionnaires et outils éducatifs pour les jeunes sourds ou malentendants, exploitant les technologies normalisées MPEG-4 et MPEG-7. Nous avons déjà orienté nos recherches en ce sens et participons aux projets européens VISICAST (IST) et PUZZLE (Socrates) actuellement en cours de développement au sein de l'Unité de Projets ARTEMIS de l'INT.

# Liste des publications associées

### Revues Internationales

T. Zaharia, F. Prêteux, "Advanced techniques for sign language indexation within the MPEG-7 framework", accepté pour publication dans *Journal of Electronic Imaging*.

### Conférences internationales avec comité de lecture et actes

T. Zaharia, F. Prêteux, "Multimedia Indexing and Retrieval: Insight into MPEG-7", *International Workshop on Multimedia Content-based Indexing and Retrieval* (MMCBIR2001), INRIA Rocquencourt, 24-25 Septembre 2001.

T. Zaharia, F. Prêteux, "Hough transform-based 3D mesh retrieval", *Proc. SPIE Conference on Vision Geometry*, San Diego, Etats-Unis, Août 2001.

T. Zaharia, F. Prêteux, "Parametric motion models for video content description within the MPEG-7 framework", *Proc. SPIE Conference on Nonlinear Image Processing and Pattern Analysis*, Vol. 4304, pp. 118-132, San Jose, Etats-Unis, Janvier 2001.

T. Zaharia, F. Prêteux, "Three-dimensional shape-based retrieval within the MPEG-7 framework", *Proc. SPIE Conference on Nonlinear Image Processing and Pattern Analysis*, Vol. 4304, pp. 133-145, San Jose, Etats-Unis, Janvier 2001.

M. Preda, T. Zaharia, F. Prêteux, "3D body animation and coding within a MPEG-4 compliant framework", *International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging* (IWSNHC3DI'99), pp. 74-78, Santorini, Grèce, 15-17 Septembre 1999.

T. Zaharia, M. Preda, F. Prêteux, "Sign language indexation within the MPEG-7 framework", *Proc. SPIE Conference on Mathematical Modeling, Bayesian Estimation and Inverse Problems*, Denver, Etats-Unis, Vol. 3816, pp. 214-228, Juillet 1999.

S. Curila, M. Curila, T. Zaharia, G. Mozelle, F. Prêteux, "A new predictive scheme for 3D mesh coding within a MPEG-4 compliant framework", *Proc. SPIE Conference on Nonlinear Image Processing X*, Vol. 4036, pp. 240-250, San Jose, Etats-Unis , Janvier 1999.

### Conférences nationales avec comité de lecture et actes

T. Zaharia, F. Prêteux, "Modèles paramétriques de mouvement pour la description des contenus vidéos dans le cadre du futur standard MPEG-7", accepté pour publication au *13ème Congrès Francophone de Reconnaissance de Forme et Intelligence Artificeille* (RFIA2002), Angers, France, Janvier 2002.

T. Zaharia, F. Prêteux, "Indexation de maillages 3D par descripteurs de forme", accepté pour publication au *13ème Congrès Francophone de Reconnaissance de Forme et Intelligence Artificeille* (RFIA2002), Angers, France, Janvier 2002.

### Rapports techniques ISO

F. Duclos, L. Fournier, F. Prêteux, T. Zaharia, "MPEG-7 mobile applications: New scenarios", *ISO/IEC JTC1/SC29/WG11, MPEG00/M6639*, La Baule, France, Octobre 2000.

T. Zaharia, F. Prêteux, "Results of 3D Shape Core Experiment", *ISO/IEC JTC1/SC29/WG11, MPEG00/M6315*, Pékin, Chine, Juillet 2000.

T. Zaharia, F. Prêteux, "The influence of the quantization step on the 3D shape spectrum descriptor performances", *ISO/IEC JTC1/SC29/WG11, MPEG00/M6316*, Pékin, Chine, Juillet 2000.

T. Zaharia, F. Prêteux, "3D Shape Core Experiment: The influence of mesh representation", *ISO/IEC JTC1/SC29/WG11, MPEG00/M6103*, Genève, Suisse, Juin 2000.

T. Zaharia, F. Prêteux, "3D Shape Core Experiment: Semantic versus geometric categorization of 3D mesh models", *ISO/IEC JTC1/SC29/WG11, MPEG00/M6104*, Genève, Suisse, Juin 2000.

T. Zaharia, F. Prêteux, "Crosscheck of the 3D shape spectrum descriptor", *ISO/IEC JTC1/SC29/WG11, MPEG00/M5917*, Noordwijkerhout, Pays Bas, Mars 2000.

T. Zaharia, F. Prêteux, "New content for the 3D shape Core Experiment: the 3D Cafe data set", *ISO/IEC JTC1/SC29/WG11, MPEG00/M5915*, Noordwijkerhout, Pays Bas, Mars 2000.

T. Zaharia, M. Preda, F. Prêteux, "3D Shape Descriptors: Results and Performance Evaluation", ISO/IEC JTC1/SC29/WG11, MPEG99/M5592, Maui, Etats-Unis, Décembre 1999.

T. Zaharia, M. Preda, F. Prêteux, "Motion Content Set: New Contributions", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5594*, Maui, Etats-Unis, Décembre 1999.

T. Zaharia, M. Preda, F. Prêteux, "Similarity Measures for Motion-based Retrieval", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5595*, Maui, Etats-Unis, Décembre 1999.

F. Prêteux, T. Zaharia, M. Preda, "Core Experiment on motion trajectories: New data set and preliminary results", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5253*, Melbourne, Australie, Octobre 1999.

T. Zaharia, F. Prêteux, M. Preda, "3D Shape spectrum descriptor", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5242*, Melbourne, Australie, Octobre 1999.

F. Prêteux, T. Zaharia, M. Preda, "Results of Core Experiment on 3D shape: The cord histogram descriptor related performances", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5240*, Melbourne, Australie, Octobre 1999.

T. Zaharia, F. Prêteux, "Results of Core Experiment on object motion: Similarity measures for parametric motion models", ISO/*IEC JTC1/SC29/WG11, MPEG99/M5237*, Melbourne, Australie, Octobre 1999.

F. Prêteux, T. Zaharia, M. Preda, "Parametric Object Motion Descriptor", *ISO/IEC JTC1/SC29/WG11*, MPEG99/M4870, Vancouver, Canada, Juillet 1999.

F. Prêteux, T. Zaharia, M. Preda, "Preliminary Results of CE on Motion Trajectory", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4871*, Vancouver, Canada, Juillet 1999.

A. Smolic, J.-R. Ohm, A.M. Tekalp, R. Mehrotra, F. Prêteux, T. Zaharia, J. Heuer, A. Kaup, "Request for Core Experiment on parametric 2-D motion descriptors in MPEG-7", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4876*, Vancouver, Canada, Juillet 1999.

F. Prêteux, S. Curila, T. Zaharia, "Hybrid scheme for geometry coding of 3D meshes: bitstream exchanges", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4900*, Vancouver, Canada, Juillet 1999.

F. Prêteux, T. Zaharia, S. Curila, "Geometry coding of 3D meshes: Integration of the polygonal and hybrid prediction methods in the 3D mesh reference software", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4553*, Séoul, Corée, Mars 1999.

F. Prêteux, M. Preda, T. Zaharia, "Predictive- versus DCT-based BAP Coding, *ISO/IEC JTC1/SC29/WG11, MPEG99/M4554*, Séoul, Corée, Mars 1999.

F. Prêteux, T. Zaharia, S. Curila, M. Curila, "Geometry Coding of 3D Meshes: Results of Core Experiment M2", *ISO/IEC JTC1/SC29/WG11, MPEG98/M4277*, Rome, Italie, Décembre 1998.

F. Prêteux, M. Preda, T. Zaharia, "Experiment on BAP coding", *ISO/IEC JTC1/SC29/WG11, MPEG98/M4283*, Rome, Italie, Décembre 1998.

F. Prêteux, M. Preda, T. Zaharia, "Preliminary results on hand BAT interpolation, *ISO/IEC JTC1/SC29/WG1*1, MPEG98/M4278, Rome, Italie, Décembre 1998.

F. Prêteux, T. Zaharia, M. Curila, S. Curila, G. Mozelle, "Geometry Compression of 3D Meshes: Results on Core Experiment M2", *ISO/IEC JTC1/SC29/*WG11, MPEG98/M4058, Atlantic City, Etats-Unis, Octobre 1998.

# Annexe 1.1 Procédure de participation aux activités de normalisation ISO/IEC/SC29

La procédure à suivre est simple et est précisée par la résolution 30 du SC29 du 22-24 Mars 1999 rappelée ci-dessous.

> SC 29 reaffirms the JTC 1 directives concerning participation at WG meetings found in the subclause 7.7.1 of the JTC 1 Directives (4th edition). In particular,
> *"Only delegates officially nominated by the NBs and the representatives of other TCs and organizations in liaison may attend WG meetings."*
> SC 29 instructs its WG Conveners to announce the above Directives at the beginning of each meeting.
> SC 29 strongly requests its National Bodies to submit delegate lists of WG meeting attendees to the SC 29 Secretariat, and the WG Convener, but not directly to the host organization, by one month prior to commencement of the meeting.
> SC 29 instructs its Secretariat to forward the lists to host organization for registration at the meeting.
> SC 29 requests the host organization to register only those persons appearing on the delegates lists, and to refer to the HoDs any persons who does not appear on the delegates lists.
> SC 29 instructs its Conveners to avide by the JTC 1 Directives and refuse attendence to WG meetings of any person who are not authorized delegates by their NBs.
> SC 29 instructs its WGs to notify their members that they should obtain Web site passwords only from HoDs of WG, not at meeting sites. These passwords should not be made public.

Petit glossaire : NB : National Body, TC : Technical Committe, HoD : Head of Delegation.

# Annexe 1.2 MPEG-4 dans ses versions successives

### *MPEG-4 V1*

Sont intégrées à MPEG-4 V1 toutes les technologies jugées suffisamment mûres et stables, c'est-à-dire la plupart des objets et outils vidéo et audio, ainsi que les outils de base pour les groupes Système, Synthétique et DMIF. MPEG-4 V1 est un standard international depuis décembre 99 et comporte notamment :

- l'animation du visage,
- le codage progressif d'images fixes par ondelettes,
- le codage de maillages 2D,

pour le groupe synthétique.

- la définition sémantique de l'interface d'abstraction du réseau, ce qui est insuffisant pour l'interopérabilité,

pour le groupe DMIF,

- les outils de base, dont le langage ECMAScript qui est une évolution de JavaScript assez sommaire,

pour le groupe Systèmes.

### *MPEG-4 V2*

En dehors des domaines de l'audio et de la vidéo naturels, MPEG-4 V2 est un standard international abouti en décembre 2000, qui est beaucoup plus sophistiqué et qui inclut :

- des outils plus adaptés au codage d'objet de forme arbitraire, tels que le *Shape Adaptive DCT* et le *Shape Adaptive Wavelet Transform*, permettant l'amélioration des performances du codage,

pour ce qui concerne le groupe Vidéo.

- l'animation du corps humain,
- le codage de maillages 3D,

pour les activités relevant du groupe Synthétique.

- la définition d'une syntaxe de signalisation permettant l'interopérabilité,

pour ce qui concerne DMIF.

- l'introduction du langage Java comme langage de script, avec la définition d'API permettant une véritable extensibilité de BIFS,
- la composition audio fondée sur un modèle perceptuel,

pour le groupe Systèmes.

### *MPEG-4 V3*

La Version 3 est clairement orientée vers l'intégration de MPEG-4 avec le World Wide Web. MPEG-4 V3 deviendra un standard international en décembre 2001. Il est donc encore possible d'intégrer des nouveaux outils dans cette version et de les faire évoluer jusqu'en décembre 2000.

Pour compléter ce bref aperçu, le lecteur est invité à consulter le document officiel [MPEG-4] spécifiant le standard 14496-1.

# Annexe 1.3 Spécification des espaces de couleur MPEG-7

- **L'espace $YC_bC_r$**
- Les relations de transformation de l'espace RGB dans l'espace $YC_bC_r$ sont linéaires et définies comme suit :

$$Y = 0.299*R + 0.587*G + 0.114*B$$
$$C_b = -0.169*R - 0.331*G + 0.500*B$$
$$C_r = 0.500*R - 0.419*G - 0.081*B$$

- **L'espace HSV** (*Hue, Saturation, Value*)

Le calcul de ses différents éléments est détaillé ci-dessous, en pseudo-code.

```
Max = max(R, G, B);
Min = min( R, G, B);
Value = max(R, G, B);
if( Max == 0 ) then
  Saturation = 0;
else
  Saturation = (Max-Min)/Max;
if( Max == Min ) // Hue is undefined (achromatic
colour);
otherwise:
if( Max == R && G > B )
  Hue = 60*(G-B)/(Max-Min)
else if( Max == R && G < B )
  Hue = 360 + 60*(G-B)/(Max-Min)
else if( G == Max )
  Hue = 60*(2.0 + (B-R)/(Max-Min))
else
  Hue = 60*(4.0 + (R-G)/(Max-Min))
```

L'espace HSV est un espace cylindrique, dont les frontières et la signification des axes sont illustrées Figure A.1.2.1.

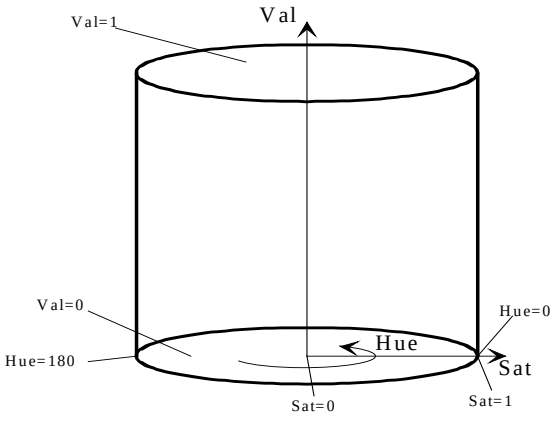



**Figure A.1.2.1.** L'espace HSV [MPEG-7Visual].

Les valeurs de V et de S sont normalisées dans l'intervalle [0, 1], tandis que la nuance H prend des valeurs dans l'intervalle [0, 360].

- **L'espace HMMD**

La signification des divers éléments intervenant dans la définition de l'espace HMMD est explicitée ci-dessous.

- Nuance (H) : exactement la même que pour l'espace HSV
- Max: Max = max(R, G, B);,
- Min: Min = min( R, G, B),
- Diff (D): indique la "pureté" de la couleur,
  Diff = Max – Min,
- Sum: donne une mesure de la luminosité de la couleur
  Sum = Max + Min.

Cet espace a l'apparence d'un double cône dont les frontières sont présentées Figure A.1.2.2.

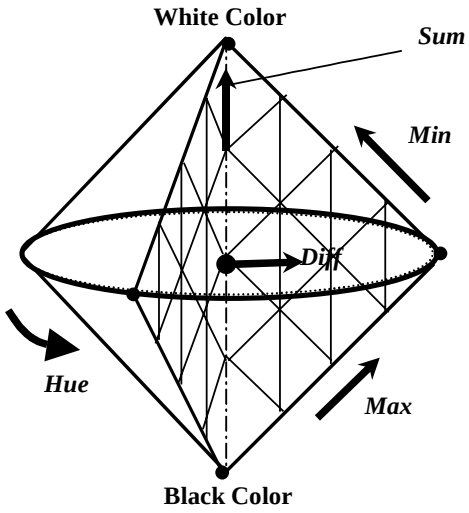



**Figure A.1.2.2.** L'espace HMMD [MPEG-7Visual].

Les valeurs des éléments diff, max, min, sum sont normalisées dans l'intervalle [0, 1], tandis que la nuance (H) peut prendre des valeurs entre 0 et 360.

# Annexe 1.4 L'algorithme LBG

Les étapes de l'algorithme LBG [Linde80] pour le calcul des couleurs dominantes sont les suivantes :

1. On cherche tout d'abord le dictionnaire composé d'un seul vecteur prototype, noté $C_0^0$. Le critère étant la minimisation de l'erreur quadratique moyenne, ce vecteur sera le centre de gravité de l'ensemble des couleurs composant l'objet.
2. On partage ce vecteur en deux vecteurs, notés $C_0^1$ et $C_1^1$, donnés par :

$$C_0^1 = C_0^0 + \varepsilon$$

$$C_1^1 = C_0^0 - \varepsilon$$

3. On classe tous les vecteurs de la base relativement à ces deux vecteurs en utilisant le critère de la distance euclidienne minimale. On calcule les centres de gravité des deux classes ainsi obtenues qui remplacerons les anciens prototypes $C_0^1$ et $C_1^1$ et on réitère le procédé jusqu'à ce que la décroissance de la distorsion moyenne devienne inférieure à un seuil pré-défini.
4. On continue la partition de chaque nouveau vecteur obtenu en deux, conformément à l'étape 2 et on réitère l'étape 3. L'algorithme s'arrête lorsque le nombre désiré de vecteurs est obtenu.

Remarquons qu'à la fin de l'algorithme LBG, il y aura une puissance entière de 2 de couleurs dominantes.

# Annexe 1.5 Principe d'incertitude et fonctions de Gabor

Soit *w(x)* une fonction réelle à valeurs complexes, telle que sa transformée de Fourier, notée $\hat{w}(\omega)$, existe et telle que les deux fonctions *w et* $\hat{w}$ soient d'énergie finie. L'inégalité de Heisenberg [Chui92, Flandrin93], connu sous le nom de principe de l'incertitude, est exprimée par l'équation suivante :

$$\Delta_w \cdot \Delta_{\hat{w}} \geq \frac{1}{2}, \tag{A1.4.1}$$

où

$$\Delta_w^2 = \frac{1}{\|w\|_2^2} \int_{-\infty}^{\infty} (x - x^*)^2 \left|w(x)\right|^2 dx \text{, avec } x^* = \frac{1}{\|w\|_2^2} \int_{-\infty}^{\infty} x \left|w(x)\right|^2 dx, \tag{A1.4.2}$$

et

$$\Delta_{\hat{w}}^2 = \frac{1}{\|\hat{w}\|_2^2} \int_{-\infty}^{\infty} (\omega - \omega^*)^2 \left|\hat{w}(\omega)\right|^2 d\omega \text{, avec } \omega^* = \frac{1}{\|\hat{w}\|_2^2} \int_{-\infty}^{\infty} \omega \left|\hat{w}(\omega)\right|^2 d\omega \tag{A1.4.3}$$

On suppose ici que la fonction *w* est choisie de telle sorte que les différentes expressions qui apparaissent ci-dessus existent.

Ici, les valeurs $x^*$ et $\omega^*$ représentent respectivement les centroïdes des fonctions $w$ et $\hat{w}$, tandis que $\Delta_w$ et $\Delta_{\hat{w}}$ désignent les largeurs de ces fonctions.

L'inégalité de Heisenberg, illustrée Figure A.1.4.1, montre qu'un signal ne peut être à la fois localisé dans le domaine spatial et fréquentiel.

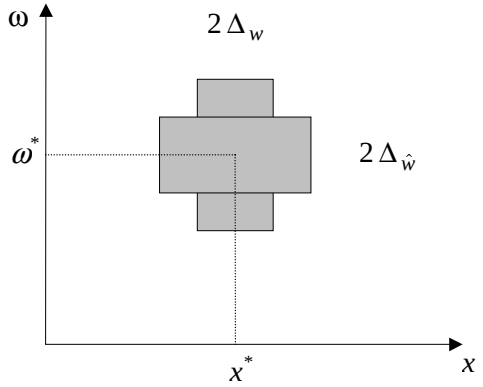



**Figure A.1.4.1.** Représentation temps-fréquence. L'aire du rectangle d'analyse doit être supérieure à 2.

En outre, on peut démontrer que les fonctions fenêtres vérifiant avec égalité l'équation (A1.4.1) sont les fonctions de Gabor, qui sont des gaussiennes complexes définies par :

$$w(x) = c e^{jax} g_\alpha(x - b) \text{, avec } g_\alpha(x) = e^{-\frac{x^2}{4\alpha}} \tag{A1.4.4}$$

où *a* et *b* prennent des valeurs réelles, α est un nombre réel positif définissant la dispersion et c est une constante de normalisation non nul.

# Bibliographie


[3DCafe] http://www.3dcafe.com.

[Adiv85] G. Adiv, “Determining three-dimensional motion and structure from optical flow generated by several moving objects”, *IEEE Trans. on PAMI*, Vol. 7, No. 4, pp. 384-401, 1985.

[Agin76] G. Agin, T. Binford, "Computer description of curved objects", *IEEE Trans. Computers*, Vol. 25, No. 4, pp. 439-449, 1976.

[AGIR] http://www.ina.fr/agir.

[Alliez01] P. Alliez, M. Desbrun, "Progressive Compression for Lossless Transmission of Triangle Meshes", *Proc. of the ACM SIGGRAPH* '01 *Conference*, 2001.

[Alshuth98] P. Alshuth, T. Hermes, L. Voigt, O. Herzog, “On video retrieval: content analysis by ImageMiner”, *Proc. of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VI*, Vol.3312, pp.236-247, Janvier 1998.

[AltaVista] www.alta-vista.com.

[ArtsVidéo] http://www.artsvideo.com/.

[Ayer95] S. Ayer, H. Sawhney, "Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding", *Proc. of the Fifth International Conference on Computer Vision* (ICCV’95), pp. 777-784, Juin 1995.

[Bailey96] R.R. Bailey, M. Srinath, "Orthogonal moment features for use with parametric and non-parametric classifiers", *IEEE Trans. on PAMI*, Vol. 18, No. 4 , pp. 389 –399, Avril 1996.

[Ballard81] D. H. Ballard, ”Generalizing the Hough transform to detect arbitrary shapes”, *Pattern Recognition*, Vol. 13, No. 2, pp. 111-122, 1981.

[Barequet97] G. Barequet, S. Kumar, "Repairing CAD models", *Proc. of IEEE Visualization'97*, pp. 363-370 , Octobre 1997.

[Becker98] J.-M. Becker, “Méthodes géométriques pour l’imagerie”, Thèse de HDR, Université Jean Monnet, St.-Etienne, Juin 1998.

[Beckman90] N. Beckman, H.-P. Kriegel, R. Schneider, B. seeger, "The r*-tree: an efficient and robust acces method for points and rectangles", *Proc. of the ACM SIGMOD*, pp. 322-331, Mai 1990.

[Benitez98] A. B. Benitez, M. Beigi, S.-F. Chang, “Using relevance feedback in content-based image metasearch**”,** *IEEE Internet Computing*, Vol. 2, No. 4, pp. 59-69, Juillet-Août 1998.

[Bennett99] N. Bennett, R. Burridge, N. Saito, “A method to detect and characterize ellipses using the Hough transform”, *IEEE Trans. on PAMI*, Vol. 21, No. 7, pp. 652 –657, Juillet 1999.

[Bezdek93] J. C. Bezdek, "Fuzzy models – what are they and why ?", *IEEE Trans. On Fuzzy Systems*, Vol. 1, No. 1, pp. 1-5, Février 1993.

[Bhandarkar94] S. M. Bhandarkar, “A fuzzy probabilistic model for the generalized Hough transform”, *IEEE Trans. On Systems, Man and Cybernetics*, Vol. 24, No. 5, pp. 745-759, Mai 1994.

[Bhattacharya00] P. Bhattacharya, H. Liu, A. Rosenfeld, S. Thompson, “Hough-transform detection of lines in 3-D space”**,** *Pattern Recognition Letters*, Vol. 21, No. 9, pp. 843-849, Août 2000.

[Biederman85] I. Biderman, “Human image understanding,: Recent research and a theory”, *Computer Vision, Graphics and Image Processing*, Vol. 32, pp. 29-73, Octobre 1985.

[Black98] M. Black, A. Jepson. “A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions”, *Proc. 5th European Conf. Computer Vision (ECCV98)*, pp. 909-924, 1998.

[Blane00] M. Blane, Z. Lei, D. B. Cooper, “The 3L Algorithm for Fitting Implicit Polynomial Curves and Surfaces to Data”, *IEEE Trans. on PAMI*, Vol. 22, No. 3; Mars 2000.

[Borgefors96] G. Borgefors, “On digital distance transforms in three dimensions”, *Computer Vision and Image Understanding*, Vol. 64, No. 3, pp. 368-376, 1996.

[Borgefors97] G. Borgefors, I. Nyström, "Efficient shape representation by minimizing the set of centres of maximal discs / spheres", *Pattern Recognition Letters*, Vol. 18, pp. 465-472, Mai 1997.

[Borgefors99a] G. Borgefors, G. Ramella, G.Sanniti di Baja, S. Svensson, "On the multiscale representation of 2D and 3D shapes", *Graphical Models and Image Processing* Vol. 61, pp. 44-62, Janvier 1999.

[Borgefors99b] G. Borgefors, I. Nystrom, G. S. D. Baja. “Computing skeletons in three dimensions”, *Pattern Recognition*, Vol. 32, pp.1225-1236, Juillet 1999.

[Bouthemy00] P. Bouthemy, M. Gelgon, F. Ganansia, "A unified approach to shot change detection and camera motion characterization”, *Rapport de recherche IRISA*, No. 1148, Novembre 2000.

[Bouthemy98] P. Bouthemy, R. Fablet, "Motion characterization from temporal cooccurrences of local motion-based measures for video indexing", *Proc. of the International. Conference. on Pattern Recognition*, ICPR'98, pp. 905-908, Août 1998.

[Bouthemy99] P. Bouthemy, M. Gelgon, F. Ganansia, "A unified approach to shot change detection and camera motion characterization”, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 7 , pp. 1030-1044, Octobre 1999.

[Bruno00] E. Bruno, D. Pellerin, “Global motion Fourier series expansion for video indexing and retrieval”, *Advances in Visual Information System, (VISUAL)*, pp. 327-337, Lyon, Novembre 2000.

[Burt83] P.J. Burt, E.H. Adelson, "The Laplacian Pyramid as a Compact Image Code" *IEEE Trans. on Communications*, Vol.COM-31, No.4, pp. 532-540, Avril 1983.

[Butlin96] G. Butlin, C. Stops, "CAD data repair", *Proc. of the 5$^{th}$ Int. Meshing Roundtable*, pp. 7-12, Octobre 1996.

[Canny86] J. Canny, "A computational approach to edge detection" *IEEE Trans. on PAMI*, Vol. 8, No. 6, pp. 679-698, 1986.

[Carey99] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, "A comparison of features for speech, music discrimination", *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 149-152, Mars 1999.

[Caspi00] Y. Caspi et M. Irani, "A step towards sequence-to-sequence alignment", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 682-689, Juin 2000.

[Catmull74] E. E. Catmull, "A Subdivision Algorithm for Computer Display of Curved Surfaces", PhD Thesis, University of Utah, Décembre 1974.

[CE-SM99.07] "Description of Core Experiments for MPEG-7 Shape/Motion descriptors", ISO/IEC JTC1/SC29/WG11, MPEG99/N2818, Juillet 1999.

[CE-CT99.12] "Description of Core Experiments for MPEG-7 Colour/Texture descriptors", ISO/IEC JTC1/SC29/WG11, MPEG99/N3090, Décembre 1999.

[Chang98] S.-F. Chang; W. Chen, H.J. Meng, H. Sundaram, D. Zhong, "A fully automated content-based video search engine supporting spatiotemporal queries", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol.8, No.5, pp.602-615, Septembre 1998.

[Cheng00] Pu-Jien Cheng; Wei-Pang Yang , "Querying video contents by motion example", *Proc. of the International Symposium on Database Applications in Non-Traditional Environments 1999 (DANTE'99)*, pp. 287 –293, 1999.

[Chuang00] J.-H. Chuang, C. H. Tsai, and M. C. Ko, "Skeletonization of three-dimensional object using generalized potential field", *IEEE Trans. on PAMI*, Vol. 22, No. 11, November 2000.

[Chui92] C. K. Chui, "Introduction to wavelets", Academic Press, Boston, MA, 1992.

[Cohen92] I. Cohen, L. Cohen, N. Ayache. "Using deformable surfaces to segment 3-d images and infer differential structures", *Computer Vision, Graphics, and Image Processing: Image Understanding*, Vol. 56, No.2, pp. 242-263, Septembre 1992.

[Cordella00] L. P. Cordella, P. Foggia, C. Sansone, M. Vento, **"**Fast graph matching for detecting CAD image components**"**, *Proc. of the 15th International Conference on Pattern Recognition* 2000, Vol. 2, pp. 1034 -1037, Septembre 2000.

[Corridoni95] J. M. Corridoni, A. Del Bimbo "Automatic video segmentation through editing analysis", *Proc. of the 8th International Conference on Image Analysis and Processing* (ICIAP '95, pp. 179-184, Springer-Verlag, Berlin, 1995.

[Corridoni98] J. M. Corridoni and A. Del Bimbo, "Structured representation and automatic indexing of movie information content"**,** *Pattern Recognition*, Vol. 31, No. 12, pp. 2027-2045, Décembre 1998.

[Curila99] S. Curila, M. Curila, T. Zaharia, G. Mozelle, F. Preteux, "A new predictive scheme for 3D mesh coding within a MPEG-4 compliant framework", *Proeedings of the. SPIE Conference on Nonlinear Image Processing X, IS&T/SPIE's Electronic Imaging* '99, San Jose, CA, Vol. 3646, pp. 240-250, Janvier 1999.

[CyberSign] Sign Language Web Site at University *Lumière* Lyon2, http://signserver.univ-lyon2.fr.

[Daubechies92] I. Daubechies, "Ten lectures on wavelets", Coll. CBMS-NSF Regional Conference Series in Applied Mathematics – 61, Philadelphie, PA, Etats-Unis, 1992.

[Davies98] C. J. Davies, M. S. Nixon, “A Hough transform for detecting the location and orientation of three-dimensional surfaces via color encoded spots”**,** *IEEE Trans. on Systems, Man and Cybernetics*, Part B, Vol. 28, No. 1, pp. 90 –95, Février 1998.

[Davis97] J. Davis, A. Bobick, “The representation and recognition of action using temporal templates”, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pp.928-934, Juin 1997.

[Davis98] J. Davis, "Mosaics of Scenes with Moving Objects*", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR98)*, pp. 354-360, Juin 1998.

[Davis99] J. Davis, "Recognizing Movement using Motion Histograms", *MIT Media Laboratory Perceptual Section Technical Report #487*, 1999.

[Deans83] S. R. Deans, “The Radon transform and some of its applications”, Wiley and Sons, New York, 1983.

[DeBerg97] M. DeBerg, “Computational geometry : algorithms and applications”, Springer-Verlag, Berlin, 1997.

[Delingette91] H. Delingette, M. Hebert, K. Ikeuchi, “Shape representation and image segmentation using deformable surfaces”, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (*CVPR’91*)*, pp. 467-472, Hawaii, Juin 1991.

[Delingette92] H. Delingette, M. Hebert, K. Ikeuchi, “Shape representation and image segmentation using deformable surfaces”, *Image and vision computing* Vol. 10, No. 3, pp. 132-144, Avril 1992.

[Delingette94] H. Delingette, “Simplex meshes: a general representation for 3D shape reconstruction”, *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1994 (CVPR '94), pp. 856 -859, Juin 1994.

[Deng98] Deng, Y., and Manjunath, B. S., "*NeTra-V: Toward an object based video representation", IEEE Trans. on Circuits and Systems for Video Technology*, Vol.8, No.5, pp. 616-627, Septembre 1998.

[DeRose98] T. DeRose. “Subdivision surface course notes”, SIGGRAPH’98 Course Notes, 1998.

[Desbrun00] M. Desbrun, M. Meyer, P. Schroeder, “Differential-Geometry Operators in nD”, MultiResolution Modelling Group, California Institute of Technology, pre-print (http://www.caltech.edu/pubs), 2000.

[Desbrun99] M. Desbrun, M. Meyer, P. Schröeder, A. H. Barr, “Implicit fairing of irregular meshes using diffusion and curvature flow”, *Proc. of the SIGGRAPH’99 Computer Graphics Conference*, Vol. 463 pp. 317-324, Août 1999.

[Deseiilligny98] M. P. Deseilligny, G. Stamon, C. Y. Suen. *"Veinerization: a new shape description for flexible skeletonization* ", *IEEE Trans. on PAMI*, Vol. 20, No. 5, pp. 505-521, Mai 1998.

[Dickinson91] S. J. Dickinson, A. Pentland, A. Rosenfeld, “From volumes to views: an approach to 3-D object recognition”, *Workshop on Directions in Automated CAD-Based Vision*, pp. 85-96, Juin 1991.

[Dickinson97] S. J. Dickinson, D. Metaxas, A. Pentland, “The role of model-based segmentation in the recovery of volumetric parts from range data**”,** *IEEE Trans. on PAMI*, Vol. 19, No. 3, pp. 259-267, Mars 1997.

[Dickinson98] S. Dickinson, A. Pentland, S. Stevenson, “Viewpoint-invariant indexing for content-based image retrieval**”,** *Proc. 1998 IEEE International Workshop on Content-Based Access of Image and Video Database*, pp. 20-30, Janvier 1998.

[DILS] "Dictionnaire illustré de la langue des signes (DILS)", Centre d´études pluridisciplinaires en langue des signes (CEPLUS), Université de Liège, Bélgique, http://www.ceplus.net/.

[Divakaran98] A. Divakaran, H. Ito, H. Sun, and T. Poon, "Scene change detection and feature extraction for indexing MPEG-2 and MPEG-4 sequences", *IEEE Trans. Circuits Syst. Video Technology* , Octobre 1998.

[Divakaran99] A. Divakaran, H. Ito, H. Sun, T. Poon, "Scene change detection and feature extraction for MPEG-4 sequences," *Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases*, San Jose, CA, January 1999.

[DoCarmo76] M. Do Carmo, "Differential geometry of curves and surfaces", Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1976.

[Dorai97] C. Dorai, A.K. Jain, "Shape Spectrum-Based View Grouping and matching of 3D Free-Form Objects", *IEEE Trans. on PAMI*, Vol. 19, No. 10, pp. 1139-1145, Octobre 1997.

[Dorai99] C. Dorai, V. Kobla “Perceived visual motion descriptors from MPEG-2 for content-based HDTV annotation and retrieval”, *Proc. of the 3rd IEEE Workshop on Multimedia Signal Processing*, pp. 147 –152, Septembre 1999.

[Dorner96] B. Dorner, "Hand shape identification and tracking for sign language interpretation", *Proc. of the 2nd International Conference on Face and Gesture recognition*, pp. 157-162, 1996.

[DucVo99] K. Duc Vo, I. Nishihara, T. Yoshida, Y. Sakai, “Precise estimation of motion vectors and its application to MPEG video retrieval”, *Proc. of the IEEE International Conference on Image Processing* (ICIP’99), Vol: 3 , pp. 279 –283, 1999.

[Duda72] R. O. Duda, P. E. Hart, “Use of the Hough transformation to detect lines and curves in pictures”, “*Communication ACM*, Vol. 15, No. 1, pp. 11-15, Janvier 1972.

[Dufaux00] F. Dufaux, J. Konrad, “Efficient, robust, and fast global motion estimation for video coding”, *IEEE Trans. on Image Processing*, Vol. 9, No. 3, pp 497 –501, Mars 2000.

[Dufaux96] F. Dufaux, F. Moscheni, “Background mosaicking for low bit rate video coding”, *Proc. of the International Conference on Image Processing* 1996, Vol. 1, pp. 673-676, Septembre 1996.

[Ekin00] A. Ekin, R. Mehrotra, and A. M. Tekalp, "Parametric description of object motion using EMUs,*" Proc. of the IEEE International Conference on Image Processing* (ICIP2000), Vol. 2, pp. 570-573, Vancouver, Canada, Septembre 2000.

[Ekin-Iso99] A. Ekin, R. Mehrota, A. M. Tekalp, “Video EMU Retrieval Using Parametric Object Motion”, *ISO/IEC JTC1/SC29/WG11, MPEG99/M5372*, Maui, Hawaii, Etats-Unis, Décembre 1999.

[Elad00] M. Elad, A. Tal, S. Ar, "Directed search in a 3D objects database using SVM", Hewlett-Packard Research Report HPL-2000-20R1, 2000.

[Fablet00-1] R. Fablet, P. Bouthemy, “Statistical motion-based object indexing using optical flow field”, *Proc. of the 15th International Conference on Pattern Recognition 2000*, Vol. 4, pp. 287-290, Septembre 2000.

[Fablet00-2] R. Fablet, P. Bouthemy, P. Pérez. “Non parametric statistical analysis of scene activity for motion-based video indexing and retrieval”, *Rapport de recherche IRISA*, No 1351, Septembre 2000.

[Fablet99] R. Fablet et P. Bouthemy, “Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval”, *Proc. 3rd Int. Conf. on Visual Information and Information Systems, VISUAL'99*, pp, 221-228, Juin 1999.

[Feder96] J. Feder, “Towards image content-based retrieval for the world wide web”, *Advanced Imaging*, Vol. 11, No. 1, pp. 26-29, Janvier 1996.

[Ferman00] A. M. Ferman, S. Krishnamachari, A. M. Tekalp, *et.al.*, "Group-of-frame/Picture Histogram Descriptors for Multimedia Applications", *Proc. of the IEEE International Conference on Image Processing* (ICIP2000), Vol. 1, pp. 65-68, Septembre 2000.

[Flandrin93] P. Flandrin, "Temps-fréquence", Coll. Traité des Nouvelles Technologies, Hermès, Paris, 1993.

[Flickner95] M. Flickner et al., "Query by image and video content: The QBIC System", *IEEE Computer*, Vol. 28, No. 9, pp. 23-32, Septembre 1995.

[Flynn89] P. J. Flynn, A. K. Jain, “On reliable curvature estimation”, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 110-116, 1989.

[Fontaine00] L. Fontaine, C. Sénac, N. Vallès-Parlaneaun, R. André-Obrecht, “Indexation de la bande sonore : les composantes parole/musique”, *XXIIIèmes Journées d'Etude sur la Parole, JEP'2000*, pp. 65-68, Juin 2000.

[Freeman94] W. T. Freeman, C.D. Weissman, "Television control by hand gestures", *Technical Report* TR-94-24, Mitsubishi Electric Research Laboratory, http://www.merl.com, 1994.

[Freeman95] W. T. Freeman, M. Roth, "Orientation histograms for hand gesture recognition, *Proc. Of the International Workshop on Automatic Face and Gesture-Recognition, IEEE Computer Society*, pp. 296-301, Zurich, Suisse, Juin 1995.

[Frey99] P. J. Frey, “Maillages. Applications aux éléments finis”, Hermès Science, Paris, 1999.

[Fröhlich95] M Fröhlich, M. Werner, "Demonstration of the interactive Graph Visualization system daVinci", *Lecture Notes in Computer Science*, No. 894, pp. 266-269, Springer-Verlag, Janvier 1995.

[Gelgon96] M. Gelgon, P. Bouthemy. "A region-level graph labeling approach to motion-based segmentation", *Rapport de recherche IRISA*, No1070, Décembre 1996.

[Gelgon98] M. Gelgon, P. Bouthemy, "Determining a structured spatio-temporal representation of video content for efficient visualization and indexing", *Proc. of the 5th European Conference on Computer Vision, ECCV'98*, Vol 1406, pp. 595-609, Springer, Freibourg, Juin 1998.

[Giese00] M. A. Giese, T. Poggio, “Morphable models for the analysis and synthesis of complex motion patterns”, *International Journal of Computer Vision*, Vol.38, No.1, pp.59-73, 2000.

[Gindikin94] S. Gindikin, “Applied problems of Radon transform”, American Mathematical Society, New York, American Mathematical Society Translations, Series 2, Vol. 162, 1994.

[Gong98] Y. Gong; G. Proietti, C. Faloutsos, "Image indexing and retrieval based on human perceptual color clustering", *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 578-583, Juin 1998.

[Graffigne98] C. Graffigne, “Stochastic modeling in image segmentation”, *Proc. of the 43rd SPIE Conference on Mathematical Modeling and Estimation Techniques in Computer Vision*, pp. 251-262, Juillet 1998.

[Greiner96] G. Greiner, “Curvature Approximation with application to surface modeling”, J. Hoschek, P. Kaklis (eds.), *Advanced Course on FAIRSHAPE*, pp. 241-252, B. G. Teubner, 1996.

[Gueziec98] A. Gueziec, G. Taubin, F. Lazarus, W. Horn, "Converting Sets of Polygons to Manifold Surfaces by Cutting and Stitching", *Proc. of the IEEE Visualization*‘98, pp. 383-390, Octobre 1998.

[Guil99] N. Guil, J. R. Cozar, E. L. Zapata, “Planar 3D object detection by using the generalized Hough transform”, *Proc.. of the 10th International Conference on Image Analysis and Processing*, pp. 358 -363, Septembre 1999.

[Günsel98] B. Günsel, A. M. Tekalp, P. J. L. van Beek, “Content-based access to video objects: Temporal Segmentation, visual summarization, and feature extraction”**,** *Signal Processing*, Vol. 66, No. 2, pp. 261-280, Avril 1998.

[Hafner95] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions" *IEEE Trans. on PAMI*, Vol. 17, No. 7, pp. 729-736, Juillet 1995.

[Hammoud00] R. Hammoud, R. Mohr, "Mixture Densities for Video Objects Recognition", *Proc. of the International Conference on Pattern Recognition*, pp. 71-75, Barcelone, Espagne, Septembre 2000.

[Hampapur94] A. Hampapur, R. Jain, T. Weymouth, “Digital video segmentation”, *Proc. oif the ACM Conference on Multimedia*, pp.357-64, Octobre 1994.

[Hampapur97] A. Hampapur, A. Gupta, B. Horowitz, C. Shu, C. Fuller, J. R. Bach, M. Gorkani, R. Jain: "Virage Video Engine", *Proc. of the SPIE Conference on Storage and Retrieval for Image and Video Databases*, pp. 188-198, Janvier 1997.

[Han00] J. Han, K.-K. Ma , “Fuzzy color histogram: an efficient color feature for image indexing and retrieval”, *Proc. of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP '00), Vol. 4 , pp. 2011-2014, Juin 2000.

[Haralick73] R. M. Haralick, K. Shanmugan, I. Dinstein, “Textural features for texture classification”, *IEEE Trans. On Systems, Man and Cybernetics*, Vol. 3, No. 6, pp. 610-621, Novembre 1973.

[Haralick79] R. Haralick, "Statistical and structural approaches to texture", *Proc. of the IEEE*, Vol. 67, No. 5, pp. 786-804, Mai 1979.

[Hebb49] D. O. Hebb, "The organization of behavior", John Wiley, 1949.

[Hebert95] M. Hebert, K. Ikeuchi, H. Delingette, “A spherical representation for recognition of free-form surfaces”, *IEEE Trans. on PAMI*, Vol. 17, No. 7, pp. 681-690, Juillet 1995.

[Hoogvorst-Iso00.03] P. Hoogvorst, “X-checking of 3-D shape descriptors technology”, *ISO/IEC JTC1/SC29/WG11, MPEG900/M5914*, Noordwijkerhout, Pays Bas, Mars 2000.

[Hoogvorst-Iso00.06] P. Hoogvorst, "Filtering of the 3D mesh models", *ISO/IEC JTC1/SC29/WG11, MPEG00/ M6101*, Genève, Suisse, Juin 2000.

[Horn81] B. K. P. Horn, B. Schunk. "Determining optical flow", Artificial Intelligence, Vol. 17, pp.185-203, 1981.

[Horn84] B. K. P Horn, "Extended Gaussian Image", *Proc. of the IEEE*, Vol. 72, pp. 1671-1686, 1984.

[Hough62] P. V. C. Hough, “Method and means for recognizing complex patterns”, U.S. Patent 3 069 654, 1962.

[Hu95] G. Hu, “3D object matching in the Hough space”**,** *IEEE International Conference on Systems, Man and Cybernetic*s, 1995, Vol. 3, pp. 2718-2723, Octobre 1995.

[Huang96] T. S. Huang, S. Mehrotra, and K. Ramchandran, "Multimedia analysis and retrieval system (MARS) project", *Proc. of the 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval*, 1996.

[Huang97] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, R. Zabih, " Image indexing using color correlograms ", *Proc. of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR’97), pp. 762-768, Juin 1997.

[Hubert81] P. J. Hubert, “Robust Statistics”, Wiley, 1981.

[Illingworth88] J. Illingworth et J. Kittler, “A survey of the Hough transform”, *Computer Vision, Graphics and Image Processing*, Vol. 44, pp. 87-116, Octobre 1988.

[Informix] http://www.informix.com.

[Irani92] M. Irani, B. Russo, S. Peleg, "Detecting and tracking multiple moving objects using temporal integration", *Proc. of the 2nd European Conference on Computer Vision (ECCV’92)*, Spinger-Verlag, pp. 282-287, Italie, Mai 1992.

[Irani97] M. Irani, H. S. Sawhney, R. Kumar, P. Anandan, “Interactive content-based video indexing and browsing”, *Proc. of the IEEE First Workshop on Multimedia Signal Processing*, pp. 313-318, Juin 1997.

[Jain89] A. K. Jain. “Fundamentals of Digital Image Processing”, Prentice Hall, Englewood Cliffs, New Jersey, 1989.

[Java3D] http://java.sun.com/products/java-media/3D/

[JDK] http://java.sun.com/products/jdk

[Jeannin00] S. Jeannin, B. Mory, "Video Motion Representation for Improved Content Access", *IEEE Trans.on Consumer Electronics*, Vol. 46, No. 3, pp. 645-655, Août 2000.

[JMF] http://java.sun.com/products/java-media/jmf/

[Johnson99] A. Johnson et M. Hebert, “Using spin-images for efficient multiple model recognition in cluttered 3-d scenes”, *IEEE Trans. on PAMI*, Vol. 21, No. 5, pp. 433-449, Mai 1999.

[Joly94] P. Aigrain, P. Joly, “The automatic real-time analysis of film editing and transition effects and its applications”, *Computers & Graphics*, Vol. 18, No. 1, pp. 93-103, Janvier-Février 1994.

[Kahn96] R. E. Kahn, M. J. Swain, P. N. Prokopowicz, R. J. Firby, "Gesture Recognition Using the Perseus Architecture", *IEEE Conference on Computer Vision and Pattern Recognition*, Juin 1996.

[Kang94] S.B. Kang, K. Ikeuchi, "The complex EGI: a new representation for 3D pose determination", *IEEE Trans. on PAMI*, vol. 16, No. 3, pp. 249-258, Mars 1994.

[Keren94] D. Keren. “Using symbolic computation to find algebraic invariants ”, *IEEE Trans. on PAMI*, Vol. 16, No. 11, pp. 1143-1149, Novembre 1994.

[Kim00] M. Kim, S. Wood and L.-T. Cheok, "Extensible MPEG-4 Textual Format (XMT)", *Proc. Of the ACM Multimedia*, Etats-Unis, Octobre-Novembre 2000, disponible en ligne à http://www.acm.org/sigs/sigmm/MM2000/ep/michelle/index.html.

[Kim99] So-Yeon Kim; Yong Man Ro, “Fast content-based MPEG video indexing using object motion”, *TENCON ’99, Proc. of the IEEE Region 10 Conference*, Vol. 2, pp. 1506-1509, 1999.

[Kiryati00] N. Kiryati, H. Kälviäinen, S. Alaoutinen, “Randomized or probabilistic Hough transform: unified performance evaluation”, *Pattern Recognition Letters*, Vol. 21, No. 13-14, pp. 1157-1164, Décembre 2000.

[Kobla97] V. Kobla, D. Doermann, K. Lin, C. Faloutsos, "Compressed domain video indexing techniques using DCT and motion vector information in MPEG video,", *Proc. of the SPIE Conference on Storage and Retrieval for Still Image and Video Databases V*, Vol. 3022, pp. 200-211, Janvier 1997.

[Koenderink86] J. Koenderink, A. Van Doorn, "Dynamic Shape", *Biological Cybernetics*, Vol. 53, No. 6, pp. 383-396, 1986.

[Koenderink90] J. Koenderink, "Solid shape", The MIT Press, Cambridge, Massachusetts, 1990.

[Koenen97] R. Koenen, F. Pereira, L. Chiariglione, “MPEG-4: Context and Objectives”, *Signal Processing: Image Communication, Special Issue on MPEG-4*, Vol. 9, No. 4, Mai 1997.

[Korn88] A. Korn, "Towards a symbolic representation of intensity changes in images", *IEEE Trans. on PAMI*, Vol. 10, No. 5, pp. 610-625, Septembre 1988.

[Krishnamachari00a] S. Krishnamachari, M. Abdel-Mottaleb, "Compact Color Descriptor for Fast Image and Video Segment Retrieval", *Proc. of the IS&T/SPIE Conference on Storage and Retrieval of Media Databases 2000*, pp. 581-589, Janvier 2000.

[Krishnamachari00b] S. Krishnamachari, A. Yamada, M. Abdel-Mottaleb, E. Kasutani, "Multimedia Content Filtering, Browsing, and Matching using MPEG-7 Compact Color Descriptors", *Proc. of the 4th International Conference on Visual Information Systems* (VISUAL2000), pp. 200-211, Novembre 2000.

[Kumar95] R. Kumar, P. Anandan, M. Irani, J. Bergen; K. Hanna, “Representation of scenes from collections of images”, *Proc. of the IEEE Workshop on Representation of Visual Scenes* 1995 (*In Conjuction with ICCV’95*), pp. 10 –17, Juin 1995

[Lefebvre83] J. Lefebvre, “Introduction aux analyses statistiques multidimensionnelles”, Masson, Paris, 1983.

[Lengagne96] R. Lengagne, P. Fua, O. Monga. “Using crest line to guide surface reconstruction from stereo”, *Proc. of the International Conference on Pattern Recognition* (ICPR'96), pp. 9-13, Vienne, Autriche, 1996.

[Li95] S. Z. Li, "Markov Random Filed modelling in Computer Vision", Computer Science Workbench, Springer-Verlag, Tokyo, 1995.

[Lienhart96] R. Lienhart, S. Pfeiffer, W. Effelsberg, “ The MoCA Workbench: support for creativity in movie content analysis”, *Proc. of the International Conference on Multimedia Computing and Systems*, pp. 314-21, Juin 1996.

[Lienhart97] R. Lienhart, W. Effelsberg, R. Jain., "VisualGREP: A Systematic Method to Compare and Retrieve Video Sequences”, *Technical Report TR-97-005*, Praktische Informatik IV, Univ. of Mannheim, Octobre 1997.

[Linde80] Y. Linde, A. Buzo, R. M. Gray, “An Algorithm for Vector Quantizer Design”, *IEEE Trans. on Communications*, Vol.COM-28, No.1, pp. 84-95, Janvier 1980.

[Liu96] F. Liu, R. Picard, "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval", *IEEE Trans. on PAMI*, Vol.18, No.7, pp.722-733, Juillet 1996.

[Loop87] C. T. Loop, "Smooth subdivision surfaces based on triangles", *Master’s Thesis*, Dep. of Mathematics, Univ. of Utah, Août 1987.

[Lounsberry94] M. Lounsberry, T. DeRose, J. Warren, "Multiresolution analysis for surfaces of arbitrary topological type", *Technical Report No. 93-10-05b*, Dept. of CS & Eng., Univ. of Washington, Janvier 1994.

[Lucas81] B. D. Lucas, T. Kanade, "An iterative image registration technique with an application to stereo vision", *Proc. of the Imaging Understanding Workshop*, pp. 121-130, 1981.

[Mahalanobis36] P.C. Mahalanobis, "On the Generalized Distance in Statistics", *Proc. of the National Institute of Science*, Vol. 12, pp 49-55, Calcutta 1936.

[Mahmoo00] T. Syeda-Mahmoo, S. Srinivasan, A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, "CueVideo: a system for cross-modal search and browse of video databases", *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 786 -787, Juin 2000.

[Malik01] J. Malik, “Shape Matching for Content based Image Retrieval”, *International Workshop on Multimedia Content-based Indexing and Retrieval (MMCBIR2001)*, INRIA Rocquencourt, 24-25 Septembre 2001.

[Mallat99] S. Mallat, “Wavelet tour of signal processing”, Academic Press, San Diego, CA, 1999.

[Manjunath96] B. S. Manjunath, W. Y. Ma, "Texture features for browsing and retrieval of image data", *IEEE-Trans. on PAMI*, Vol.18, No.8, pp.837-842, Août 1996.

[Marr76] D. Marr, "Early processing of visual information", *Proc. of the Royal Society of London*, B275, pp. 483-519, 1976.

[Marr82] D. Marr, "Vision", Freeman, San Francisco, 1982.

[Matsuo94] H. Matsuo, A. Iwata, "3D Object Recognition using MEGI model from range data", *Proc. of the 12$^{th}$ International Conference on pattern Recognition*, pp. 843-846, 1994.

[Mazière00] M. Maziere, F. Chassaing, L. Garrido, P. Salembier, “Segmentation and tracking of video objects for a content-based video indexing context”, *Proc. of the IEEE International Conference on Multimedia and Expo, 2000 (ICME 2000)*, Vol. 2, pp. 1191-1194, Août 2000.

[McIvor97] A.M. McIvor et R.J. Valkenburg, “A comparison of local surface geometry estimation method”, *Machine Vision and Applications* 10(1), Springer-Verlag, Janvier 1997.

[McLaughlin98] R. A. McLaughlin, “Randomized Hough Transform: Improved ellipse detection with comparison**”,** *Pattern Recognition Letters*, Vol. 19, No. 3-4, pp. 299-305, Mars 1998.

[Medioni00] G. Medioni, A. François, “3D sructures for generic object recognition”, *Proc. of the 15th International Conference on Pattern Recognition*, Vol. 1, pp. 30-37, Septembre 2000.

[Mokhtarian92] F. Mokhtarian, A. K. Mackworth, "A theory of multiscale, curvature-based shape representation for planar curves", *IEEE Trans. on PAMI*, Vol. 14, No. 8, pp. 789-805, Août 1992.

[Mosheni96] F. Moscheni, F. Dufaux, M. Kunt, “Object tracking based on temporal and spatial information”**,** *Proc. of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP-96), Vol. 4 , pp. 1914 –1917, Mai 1996.

[Mozelle98] G. Mozelle, F. Prêteux, J.E. Viallet, "Tele-sign : A compression framework for sign language distant communication", *Proc. SPIE Conference on Mathematical Modeling and Estimation Techniques in Computer Vision*, Vol. 3457, pp. 94-110, San Diego, Etats-Unis, July 1998.

[Mozelle98T] G. Mozelle, "Compression de données multiples et dynamiques", Thèse de Doctorat de l’Université Paris V, UFR de Mathématiques-Informatique, Juillet 1998.

[MPEG-4] Standard International *ISO/ IEC 14496,"*Information technology - Coding of audio-visual objects".

[MPEG-7Appl] Anthony Vetro (Editor), “MPEG-7 Applications Document v.10”, *ISO/IEC JTC1/SC29/WG11, MPEG01/N3934*, Janvier 2001.

[MPEG-7-Audio] Text of ISO/IEC 15938-4/FDIS Information technology - Multimedia content description interface – Part 4 Audio”, *ISO/IEC JTC1/SC29/WG11, MPEG01/N4004*, Mars 2001.

[MPEG-4-AudioCfP] “Call For Proposals for New Tools for Audio Coding”, *ISO/IEC JTC1/SC29/WG11 MPEG01/N3794*, Janvier 2001.

[MPEG-4-VideoCfP] "Call for Proposals for new tools to further improve video coding efficiency", *ISO/IEC JTC1/SC29/WG11 MPEG00/N3671*, Octobre 2000.

[MPEG-7CfP] WG11 Requirements Group, “Guide to submitting to the MPEG-7 CfP”, *ISO/IEC JTC1/SC29/WG11, MPEG98/N2568*, Rome, Italie, Décembre 1998.

[MPEG-7Content] WG11 Requirements Group, “Guide to obtaining the MPEG-7 Content Set”, *ISO/IEC JTC1/SC29/WG11, MPEG98/N2570*, Rome, Italie, Décembre 1998.

[MPEG-7Eval] WG11 Requirements Group, “MPEG-7 Evaluation Guide”, *ISO/IEC JTC1/SC29/WG11, MPEG98/N2569*, Rome, Italie, Décembre 1998.

[MPEG-7MDS] P. van Beek, A. B. Benitez, J. Heuer, J. Martinez, P. Salembier, Y. Shibata, J. R. Smith, T. Walker (Editors), "Text of ISO/IEC 15938-5 FDIS Information Technology -

Multimedia Content Description Interface - Part 5 Multimedia Description Schemes", *ISO/IEC JTC1/SC29/WG11, MPEG01/N4205*, Juillet 2001.

[MPEG-7REQ] F.Pereira (Editor), "MPEG-7 Requirements Document V.15", *ISO/IEC JTC1/SC29/WG11, MPEG01/N4320*, Juillet 2001.

[MPEG-7-Visual] L. Cieplinski, W.-Y. Kim, J.-R. Ohm, M. Pickering, A. Yamada, "Text of ISO/IEC 15938-3/FDIS Information technology - Multimedia content description interface – Part 3 Visual", *ISO/IEC JTC1/SC29/WG11, MPEG01/N4358*, Juillet 2001.

[MPEG-7XM] A.Yamada, M. Pickering, S. Jeannin, L. Cieplinski, J. R. Ohm, M. Kim (Editeurs), " Text of ISO/IEC FCD 15938-3 Information technology - Multimedia content description interface: Visual ", *ISO/ IEC JTC1/ SC29/ N4063*, Singapore, Mars 2001.

[MPEG-Press07.01] Convenor of MPEG, "MPEG Press Release", *ISO/IEC JTC1/SC29/WG11 MPEG01/N4140*, Juillet 2001.

[MPEG-Struct] http://mpeg.telecomitalialab.com/terms_of_reference.htm.

[Murali97] T. M. Murali, T. A. Funkhouser, "Consistent solid and boundary representations", *1997 SIGGRAPH Symposium on Interactive 3D Graphics*, pp. 155-162, Mars 1997.

[Murao-Iso99.02] T. Murao, "Descriptors of polyhedral data for 3D-shape similarity search", *Proposal P177, MPEG-7 Proposal Evaluation Meeting*, Lancaster, UK, Février 1999.

[Murao-Iso99.07] T. Murao, E. Paquet "Core Experiment Proposal on 3D shape for MPEG-7", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4679*, Juillet 1999.

[Murao-Iso99.10] T. Murao, E. Paquet "A Report of Results in 3D Shape Descriptor Core Experiment Stage One", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5021*, Melbourne, Australia, Octobre 1999.

[Nagasaka92] A. Nagasaka, Y. Tanaka "Automatic video indexing and full search for objetc appearances", *IFIP Trans., Visual Database Systems II*, Vol. 1, No. 7, pp. 113-128, 1992.

[Nastar96] "Generalized Image matching : Statistical learning of physically-based deformations", *Research Report MIT Media Laboratory*, No. 368, Avril 1996.

[Nastar97] C. Nastar, "The Image Shape Spectrum for Image Retrieval", *INRIA Technical Report RR-3206*, Juillet 1997.

[Niblack93] W. Niblack et al., "The QBIC Project: Querying Images by Content Using Color, Texture and Shape", *Proc. of the SPIE Conference on Storage and Retrieval for Image and video Databases III*, Vol. 2, pp. 173-87, Février 1993.

[NRC92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, "Numerical Recipes in C", Cambridge Univ. Press, 1992.

[Odobez95]. J.-M. Odobez, P. Bouthemy, "Robust multiresolution estimation of parametric motion models", *Journal of Visual Communication and Image Representation*, Vol. 6, No. 4, pp. 348-365, Décembre 1995.

[Odobez98] J.-M. Odobez, P. Bouthemy, "Direct incremental model-based image motion segmentation for video analysis", *EURASIP, Signal Processing*, Vol. 66, No. 2, pp.143-156, Avril 1998.

[Oracle] http://www.oracle.com.

[Ortega00] M. Ortega, S. Mehrotra, K. Chakrabarti, K. Porkaew, "WebMARS: A Multimedia Search Engine", *Proc. of the SPIE Conference on Electronic Imaging 2000: Internet Imaging*, pp. 314-321, Janvier, 2000.

[Osada01] R. Osada, T. Funkhouser, B. Chazelle, D.P. Dobkin, "Matching 3D models with shape distributions", *Proc. of the International Conference on Shape Modeling and Applications*, pp. 154-166, Mai 2001.

[Pao92] D.C.W. Pao; H.F. Li, R. Jayakumar, "Shapes recognition using the straight line Hough transform: theory and generalization", *IEEE Trans. on PAMI*, Vol., 14 No. 11, pp. 1076-1089, Novembre 1992.

[Paquet00] E. Paquet, M. Rioux, A. Murching, T. Naveen, A. Tabatabai, "Description of shape information for 2D and 3D objects", *Signal Processing: Image Communications*, Vol. 16, pp. 103-122, Septembre 2000.

[Paquet-Iso99] E. Paquet, *Proposals P007-010, MPEG-7 Proposal Evaluation Meeting*, Février 1999.

[Pass96] G. Pass, R. Zabih, J. Miller, "Comparing images using color coherence vectors Comparing images using color coherence vectors ", *Proc. of ACM Multimedia'96*, pp. 65-73, Boston, Novembre 1996.

[Pentland91] A. Pentland, S. Sclaroff, "Closed form solutions for physically-based shape modelling and recognition", *IEEE Trans. on PAMI*, Vol. 13, No. 7, pp. 715-729, Juillet 1991.

[Pereira00] F. Pereira, "MPEG-4: concepts, tools and applications", *Journal de Réseaux et systèmes répartis - calculateurs parallèles, Special Issue on Image and Video*, Vol. 12, No. 3-4, pp. 299-313, 2000.

[Persoon77] E. Persoon, K. Fu, "Shape discrimination using Fourier descriptors", *IEEE Trans. on. Systems, Man, and Cybernetics*, Vol. 7, No. 3, pp. 170-178, Mars 1977.

[Philip94] K. P. Philip, E. L. Love, D. D. McPherson, N. L. Gotteiner, W. Stanford, K. B. Chandran, "The fuzzy Hough transform-feature extraction in medical images", *IEEE Trans. on Medical Image Processing*, Vol. 13, No. 2, Juin 1994.

[Pope98]. A. Pope, R. Kumar, H. Sawhney, C. Wan, "Video abstraction: summarizing video content for retrieval and visualization", *Proc. of the Thirty-Second Asilomar Conference on Signals, Systems & Computers*, Vol. 1, pp. 915-919, Novembre 1998.

[Porkaew99] K. Porkaew, K. Chakrabarti, S. Mehrotra, "Query refinement for multimedia similarity retrieval in MARS", *Proc. of the ACM Multimedia Conference*, pp. 235-238, Novembre 1999.

[Poularakis00] A. Poularakis (Edt.), "The transforms and applications handbook", Electrical Engineering Handbook Series, Boca Raton, 2000.

[Pratt78] W. K. Pratt, "Digital Image Processing", John Wiley, New York, 1978.

[Preda99] M. Preda, T. Zaharia, F. Prêteux, "3D body animation and coding within a MPEG-4 compliant framework", *International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging* (IWSNHC3DI'99), pp. 74-78, Santorini, Grèce, Septembre 1999.

[Prêteux92] F. Prêteux, "On a distance function approach for gray-llevel mathematical morphology", Mathematical Morphology in Image Processing, Edited by E. R. Dougherty, pp. 323-350, Dekker, 1992.

[Prêteux-Iso98.10] F. Prêteux, T. Zaharia, M. Curila, S. Curila, G. Mozelle, "Geometry Compression of 3D Meshes: Results on Core Experiment M2", ISO/IEC JTC1/SC29/WG11, MPEG98/M4058, Octobre 1998.

[Prêteux-Iso98.12] F. Prêteux, T. Zaharia, S. Curila, M. Curila, "Geometry Coding of 3D Meshes: Results of Core Experiment M2", ISO/IEC JTC1/SC29/WG11, MPEG98/M4277, Decembre 1998.

[Prêteux-Iso99.03] F. Prêteux, T. Zaharia, S. Curila, "Geometry coding of 3D meshes: Integration of the polygonal and hybrid prediction methods in the 3D mesh reference software", ISO/IEC JTC1/SC29/WG11, MPEG99/M4553, Mars1999.

[Prêteux-Iso99.06m] F. Prêteux, T. Zaharia, M. Preda, "Parametric Object Motion Descriptor", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4870*, Vancouver, BC, Canada, Juillet 1999.

[Prêteux-Iso99.07] F. Prêteux, S. Curila, T. Zaharia, "Hybrid scheme for geometry coding of 3D meshes: bitstream exchanges", ISO/IEC JTC1/SC29/WG11, MPEG99/M4900, Juillet 1999.

[Preteux-Iso99t] F. Prêteux, T. Zaharia, M. Preda, "Preliminary Results of CE on Motion Trajectory", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4871*, Vancouver, Juillet 1999.

[Queck95] F. Quek, "Eyes in the interface", *International Journal of Vision and Image Computing*, Vol. 13, pp. 511-525, Août 1995.

[Quek94] F. Quek, "Towards a vision based hand gesture interface", *Virtual Reality and Software Technology Conference*, pp. 17-29, Singapore, 1994.

[RandomSets1] G. Matheron, “Random Sets and Integral Geometry”, Wiley, New York, 1975.

[Rehg93] J.M. Rehg, T. Kanade, "Digit Eyes: vision-based human hand tracking", Carnegie Mellon University, School of Computer Science Technical Report, CMU-CS-93-220, Décembre 1993.

[RNRT] http://www.telecom.gouv.fr/rnrt/

[Rui97-Iso.a] Y. Rui, T. S. Huang, S. Mehrotra, "MARS and Its Applications to MPEG-7", *ISO/IEC JTC1/SC29/WG11 MPEG97/M2900*, Juillet 1997.

[Rui97-Iso.b] Y. Rui, T. S. Huang, S. Mehrotra, "Suggestions to the Draft of MPEG-7 Requirements", *ISO/IEC JTC1/SC29/WG11 M3107*, MPEG97, 1997.

[Rui98] Y. Rui, T. S. Huang, M. Ortega, S. Mehrotra, “Relevance feedback: A power tool in interactive content-based image retrieval”, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 8, No. 5, pp. 644-655, Septembre 1998.

[Sahouria98] E.Sahouria,; A. Zakhor, “Content Analysis of Video Using Principal Components”**,** *Proc. of the International Conference on Image Processing (ICIP’98)*, pp. 541 –545, 1998.

[Sahouria99] E.Sahouria, A. Zakhor, “Content Analysis of Video Using Principal Components”**,** *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 8, pp. 1290-1298, Décembre 1999.

[Salembier99] P. Salembier, F. Marques, “Region-based representations of image and video: segmentation tools for multimedia services**”,** *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 8, pp. 1147-1169, Décembre 1999

[Sawhney95] *H.S. Sawhney, S. Ayer, M. Gorkani.*, "Model-based 2D&3D dominant motion estimation for mosaicing and video representation", *Proc. Fifth International Conference on Computer Vision* 1995, pp. 583-590, 1995.

[Schlenzig94] J. Schlenzig, E. Hunter, R. Jain, "Recursive identification of gesture inputs using hidden Markov models", *Proc. of the 2nd Annual Conference on Applications of Computer Vision*, pp. 187-194, Décembre 1994.

[Sclaroff94] S. Slaroff, A. Pentland, "Physically-based combinations of views : Representing rigid and non-rigid motion", *Proc. of the IEEE Workshop on Nonrigid and Articulate Motion*, Austin, Novembre 1994.

[Sclaroff97] S. Sclaroff, "Deformable prototypes for encoding shape categories in image databases", *Pattern Recognition*, Vol. 30, No. 4, pp. 627-641, Avril 1997.

[Seamless] http://www.seamless-solutions.com.

[Serra82] J. Serra, "Image Analysis and Mathematical Morphology", New York: Academic, 1982.

[Serra88] J. Serra, "Image Analysis and Mathematical Morphology. Volume 2: Theoretical Advances", Academic Press, London, 1988.

[Sheehy96] D. J. Sheehy, C. G. Armstrong, and D. J. Robinson, "Shape description by medial surface construction", *IEEE Trans. on Visualization and Computer Graphics*, Vol. 2, No. 1, pp. 62-72, March 1996.

[SM-CE99.07] "Description of Core Experiments for MPEG-7 Shape/Motion descriptors**",** ISO/IEC JTC1/SC29/WG11, MPEG99/N2818, Sydney, Australie, Juillet 2001.

[SMIL] http://www.w3.org/TR/smil20.

[Smith96] J. R. Smith and S.-F. Chang, "VisualSEEk: a fully automated content-based image query system", *Proc. of the ACM Multimedia*, pp. 87-98, Novembre 1996.

[Smolic99] A. Smolic, T. Sikora, J.-R. Ohm, "Long-term global motion estimation and its application for sprite coding, content description, and segmentation", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 8, pp. 1227-1242, Décembre1999.

[Smolic-Iso99a] A. Smolic, J.-R. Ohm, A. M. Tekalp, R. Mehrotra, F. Preteux, T. Zaharia, J. Heuer, A. Kaup, "Request for Core Experiment on parametric 2-D motion descriptors in MPEG-7", *ISO/IEC JTC1/SC29/WG11, MPEG99/M4876*, Vancouver, Canada, Juillet 1999.

[Smolic-Iso99b] A. Smolic, M. Hoeynck, J.-R. Ohm, "Results of MPEG-7 Core Experiment on Parametric Object Motion", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5394*, Maui, Hawaii, Etats-Unis, Décembre 1999.

[Soodamani98] R. Soodamani, *Z.Q.* Liu, "A novel fuzzy Hough transform for shape representation", *Proc. of the 1998 IEEE International Conference on Fuzzy Systems*, Vol 2, pp. 1605-1608, Mai 1998.

[Spivak79] M. Spivak, "A comprehensive introduction to differential geometry", 2nd ed., Houston: Publish or Perish, 1979.

[SQL] International Standard, ISO/IEC 9075:1992, "Information Technology - Database Languages - SQL".

[Starner95] T. Starner, A. Pentland, "Real-time American Sign Language recognition from video using hidden Markov models", *Proc. of the IEEE International Symposium on Computer Vision*, pp. 265-270, Comput. Soc. Press, Los Alamitos, CA, Etats-Unis, Novembre 1995.

[Starner96] T. Starner, A. Pentland. "Real-time ASL recognition from video using HMM's", *Rapport de recherche MIT Media Laboratory no. 375*, 1996.

[Starner98] T. Starner, J. Weaver, A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video", *IEEE-Trans. on PAMI*, Vol.20, No.12, pp.1371-1375, Décembre 1998.

[Stein98] G. P. Stein, "Tracking from multiple view points: Self calibration of space and time", DARPA IU Workshop, pp. 1037-1042, 1998.

[Stiller99] C. Stiller, J. Konrad, "Estimating motion in image sequences", *IEEE Signal Processing Magazine*, Vol. 16, No. 4, pp. 70-91, Juillet 1999.

[Stricker94] M. Stricker, M. Swain, "The capacity of color histogram indexing", *Proc. of the 1994 IEEE Conference on Computer Vision and Pattern Recognitio*n (CVPR'94), pp. 704-708, Juin 1994.

[Subrahmonia96] J. Subrahmonia, D. B. Cooper, and D. Keren, "Practical, Reliable, Bayesian Recognition of 2D and 3D Objects Using Implicit Polynomials and Algebraic Invariants", *IEEE Trans. on PAMI*, Vol. 18, No. 5, pp. 505-519, Mai 1996.

[Swain91] M. Swain, D. Ballard, "Color Indexing", *International Journal of Computer Vision*, Vol. 7, No. 11, pp. 11-32, Novembre 1991.

[Szeliski96] R. Szeliski, H.-Y. Shum, "Motion Estimation with quadtree splines", *IEEE Trans. On PAMI*, Vol. 18, No. 12, pp. 1199-1210, Décembre 1996.

[Tamura78] H. Tamura, S. Mori, T. Yamawaki, "Texture features corresponding to visual perception", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 8, No. 6, pp. 460-473, Juin 1978.

[Tang99] C. K. Tang, G. Medioni, "Robust Estimation of Curvature Information for Shape Description", *Proc. of the International Conference on Computer Vision* (ICCV'99), Vol.1, pp. 426-433, Septembre 1999.

[Taubin91] G. Taubin, D. B. Cooper, "Recognition and positioning of rigid objects using algebraic moment invariants", *Geometric Methods in Computer Vision*, SPIE'91, Vol. 1570, pp. 1284-1288, 1991.

[Taubin92] G. Taubin et D.B. Cooper, "Object recognition based on moment (or algebraic) invariants", *J.L. Mundy and A. Zisserman, editors, Geometric Invariants in Computer Vision*, pp. 375-397, MIT Press, 1992.

[Taubin95] G. Taubin, "Estimating the tensor of curvature of a surface from a polyhedral approximation", *Proc. of the International Conference on Computer Vision*, pp. 902-907, Cambridge, MA, Juin 1995.

[Taubin95fair] G. Taubin, "A Signal Processing Approach to Fair Surface Design", *Proc. of the Computer Graphics Conference* SIGGRAPH'95, pp. 351-358, Août 1995.

[Taubin98] G. Taubin, W.P. Horn, J. Rossignac, F. Lazarus, "Geometry coding and VRML", *Proeedings. of the IEEE, Special issue on Multimedia Signal Processing*, Vol. 86, No. 6, pp. 1228-1243, Juin 1998.

[Teague80] M. R. Teague, "Image analysis via the general theory of moments", *Journal of the Optical Society of America*, Vol. 70, pp. 920-930, Août 1980.

[Terzopoulos91] D. Terzopoulos, D. Metaxas, "Dynamic 3D models with local and global deformations: Deformable superquadrics", *IEEE Trans. on PAMI*, Vol. 13, No. 7, pp. 703-714, Juillet 1991.

[Tolias99] Y. Tolias, S. Panas, L. H. Tsoukalas, "FSMIQ: fuzzy similarity matching for image queries", *Proc. of the 1999 International Conference on Information Intelligence and Systems*, pp. 249 –254,1999.

[Truchetet98] F. Truchetet, "Ondelettes pour le signal numérique", Hermès, Paris, 1998.

[Tsuji78] S. Tsuji et F. Matsumoto, , "Detection of ellipses by a modified Hough transform", *IEEE Trans. on Computers*, Vol. C-27, No.8, pp.777-781, Août 1978.

[Turk91] M. A. Turk, A. P. Pentland, "Face recognition using eigenfaces", *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'91), pp. 586-591, Juin 1991.

[Ullmann 76] J. R. Ullmann. "An algorithm for subgraph isomorphism, *Journal of the ACM*, Vol. 1, No. 23, pp. 31-42, Janvier 1976.

[Ulupinar93] F. Ulupinar, R. Nevatia, "Perception of 3D surfaces from 2D contours", *IEEE Trans. on PAMI*, Vol. 15, No. 1, pp. 3-18, Janvier 1993.

[UML] *Rational Software*, http://www.rational.com.

[Vertan00a] C. Vertan, N. Boujemaa, "Embedding Fuzzy Logic in Content Based Image Retrieval", *Proc. of the 19th International Meeting of the North American Fuzzy Information Processing Society* (NAFIPS 2000), pp. 85-89, Juillet 2000.

[Vertan00b] C. Vertan, V. Buzuloiu, "Fuzzy nonlinear filtering of color images : a survey", E. Kerre, M. Nachtegael (Edts.), Fuzzy Techniques in Image Processing, Physica Verlag, Heidelberg, Allemagne, 2000.

[Video-Q] http://www.ctr.columbia.edu/VideoQ/visual.html.

[Visicast] Projet européen Visicast, www.visicast.co.uk.

[VRML97] International standard "The Virtual Reality Modeling Language", *ISO/ IEC 14772-1*, 1997.

[Wactlar00] H. D. Wactlar, "Informedia - Search and Summarization in the Video Medium", *Proc. of Imagina 2000 Conference*, Monaco, Janvier-Février 2000.

[Wang90] Y P. Wang, T. Pavlidis, "Optimal correspondence of string subsequences **",** *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 11, pp. 1080-1087, Novembre 1990.

[Wang94] J. Y. A Wang, E. H Adelson, "Representing moving images with layers**"**, *IEEE Trans. on Image Processing*, Vol. 3, No. 5, pp. 625-638, Septembre 1994.

[Wu00] P. Wu, B. S. Manjunath, S. Newsam, H. D. Shin, "A texture descriptor for browsing and similarity retrieval", *Signal Processing: Image Communication*, Vol.16, No.1-2, pp.33-43, Septembre 2000.

[Wu93] Wu, K., M. D. Levine, “3D object representation using parametric geons”, *Research Report TR-CIM-93-13, McGill Research Centre for Intelligent Machines*, McGill University, Montreal, Canada, 1993.

[Wu94] K. Wu, M.D. Levine, “Recovering of parametric geons from multiview range data”, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 159-166, Juin 1994.

[X3D] http://www.web3d.org/x3d.html

[Xerces] http://xml.apache.org/

[Xin96] Y. Xin, “Geometry of harmonic maps”, Birkhauser, 1996.

[XMLSchema] http://www.w3.org/XML/Schema

[Xu90] L. Xu, E. Oja, P. Kultanen, “A new curve detection method: randomized Hough transform (RHT)”, *Pattern recognition Letters*, Vol. 11, No. 5, pp. 331-338, Mai 1990.

[Yahoo] www.yahoo.com.

[Yamada00] A. Yamada, E. Kasutani, M. Ohta, K. Ochiai, H. Matoba, "Visual program navigation system based on spatial distribution of color", *Proc. of the IEEE International Conference on Consumer Electronics*, pp. 280-281, Juin 2000.

[Yang99] M.-H. Yang, N. Ahuja, "Recognizing Hand Gestures Using Motion Trajectories", *Proc. of the 1999 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 99)*, Vol. 1, pp. 466-472, Juin, 1999.

[Yeung97] M. Yeung, B. Yeo, "Time constrained clustering for segmentation of the video into story units", Proc. of the *IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases*, Vol. 3022, pp. 45-58, Février 1997.

[You01] W. You, K. W. Lee, J.-G. Kim, J. Kim, O.-S. Kwon, “Content-based video retrieval by indexing object's motion trajectory”, *Proc. of the International Conference on Consumer Electronics 2001 (ICCE 2001)*, pp. 352-353, Juin 2001.

[Yu01] T. Yu, Y. Zhang, “Retrieval of video clips using global motion information”**,** *Electronics Letters* , Vol. 37, No. 14, pp. 893-895, Juillet 2001.

[Zabih95 ] R. Zabih, J. Miller, K. Mai, "A Feature-Based Algorithm for Detecting and Classifying Scene Breaks", *Proc.of the ACM Multimedia 95*, pp. 189-200, Novembre 1995.

[Zaharia01] T. Zaharia, F. Prêteux, "Parametric motion models for video content description within the MPEG-7 framework", *Proc. of the SPIE Conference on Nonlinear Image Processing and Pattern Analysis*, San Jose, Etats-Unis, 22-23 Janvier 2001.

[Zaharia99] T. Zaharia, M. Preda, F. Prêteux, "Sign language indexation within the MPEG-7 framework", *Proc. SPIE Conference on Mathematical Modeling, Bayesian Estimation and Inverse Problems*, Denver, Etats-Unis, Vol. 3816, pp. 214-228, Juillet 1999.

[Zaharia99] T. Zaharia, M. Preda, F. Prêteux. "Sign Language Indexation within the MPEG-7 Framework", *Proc. of the SPIE Conference on Mathematical Modeling, Bayesian Estimation, and Inverse Problems*, Vol. 3816, Denver, Juillet 1999.

[Zaharia-Iso00.03a] T. Zaharia, F. Prêteux, “ New content for the 3D Shape Core Experiment: the 3D Cafe data set”, *ISO/IEC JTC1/SC29/WG11, MPEG00/M5915*, Noordwijkerhout, Pays Bas, Mars 2000.

[Zaharia-Iso00.03b] T. Zaharia, F. Prêteux, “Crosscheck of the 3D shape spectrum descriptor”, *ISO/IEC JTC1/SC29/WG11, MPEG00/M5917*, Noordwijkerhout, Pays Bas, Mars 2000.

[Zaharia-Iso00.06a] T. Zaharia, F. Prêteux, "3D Shape Core Experiment: Semantic versus geometric categorization of 3D mesh models", *ISO/IEC JTC1/SC29/WG11, MPEG00/M6104*, Genève, Suisse, Juin 2000.

[Zaharia-Iso00.06b] T. Zaharia, F. Prêteux, "3D Shape Core Experiment: The influence of mesh representation", *ISO/IEC JTC1/SC29/WG11, MPEG00/M6103*, Genève, Suisse, Juin 2000.

[Zaharia-Iso00.07a] T. Zaharia, F. Prêteux, “Results of 3D Shape Core Experiment”, *ISO/IEC JTC1/SC29/WG11, MPEG00/M6315*, Pékin, Chine, Juillet 2000.

[Zaharia-Iso00.07b] T. Zaharia, F. Prêteux, "The influence of the quantization step on the 3D shape spectrum descriptor performances", ISO*/IEC JTC1/SC29/WG11, MPEG00/M6316*, Pékin, Chine, Juillet 2000.

[Zaharia-Iso99.02] T. Zaharia, F. Prêteux, "Motion descriptor: perspective transformation parameters and object trajectory", *Proposal P351, MPEG-7 Proposal Evaluation Meeting*, Lancaster, UK, Fevrier 1999.

[Zaharia-Iso99.10s] T. Zaharia, F. Prêteux, M. Preda, "3D Shape spectrum descriptor", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5242* , Melbourne, Australie, Octobre 1999.

[Zaharia-Iso99.12s] T. Zaharia, M. Preda, F. Prêteux, “3D Shape Descriptors: Results and Performance Evaluation”, *ISO/IEC JTC1/SC29/WG11, MPEG99/M5592*, Maui, Etats-Unis, Décembre 1999.

[Zaharia-Iso99c] T. Zaharia, F. Prêteux, "Results of Core Experiment on object motion: Similarity measures for parametric motion models", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5237* , Melbourne, Australie, Octobre 1999.

[Zaharia-Iso99d] T. Zaharia, M. Preda, F. Prêteux, "Similarity Measures for Motion-based Retrieval", *ISO/IEC JTC1/SC29/WG11, MPEG99/M5595*, Maui, Hawaii, Etats-Unis, Décembre 1999.

[Zaharia-Iso99e] T. Zaharia, M. Preda, F. Prêteux, "Motion Content Set: New Contributions”, *ISO/IEC JTC1/SC29/WG11, MPEG99/M5594*, Maui, Hawaii, Etats-Unis, Décembre 1999.

[Zahn72] C. T. Zahn, R. Z. Roskies, "Fourier descriptors for plane closed curves," *IEEE Trans. on Computers*, Vol.C-21, No.3, pp. 269-281, Mars 1972.

[Zerroug94] M. Zerroug, R. Nevatia, “Segmentation and recovery of SHGC from a real intensity image”, *Proc. of the 3rd Conference on Computer Vision, Lecture notes in Computer Science*, pp. 319-330, Springer-Verlag, Mai 1994.

[Zhang01] C. Zhang, T. Chen, “Efficient feature extraction for 2D/3D objects in mesh representation”, *Proc. of the International Conference on Image Processing* (ICIP 2001), 2001.

[Zhang94] J. Zhang, "Region-based road recognition for guiding autonomous vehicles", PhD Thesis, Department of Computer Science, University of Karlsruhe, Allemagne, 1994.

[Zhang99] D. Zhang and M. Hebert, "Harmonic maps and their applications in surface matching" *Proeedings of the International Conference on Computer Vision and Pattern recognition* (CVPR '99), pp. 524-530, Juin 1999.

[Zhou99] Y. Zhou, and A. W. Toga, “Efficient skeletonization of volumetric objects”, *IEEE Trans. on Visualization and Compuer Graphics*, Vol. 5, No. 3, pp. 196-209, Juillet-Septembre 1999.

[Zhou00] W. Zhou, A. Vellaikal, C. J. Kuo, “Video analysis and classification for MPEG-7 applications”, *Proc. of the International Conference on Consumer Electronics*, pp. 344-345, Juin 2000.

[Zhou-JY00] J. Y. Zhou, E. P. Ong, C. C. Ko, "Video object segmentation and tracking for content-based video coding", *Proc. of the IEEE International Conference on multimedia and Expo (ICME2000)*,Vol. 3, pp. 1555-1558, 2000.

[Zusne70] L. Zusne, "Contemporary theory of visual form perception: III. The global theories", Academic Press, 1970.