ETUDE DE LIGNES D'INTERET NATURELLES POUR LA REPRESENTATION D'OBJETS EN VISION PAR ORDINATEUR

Thi-Thanh-Hai Tran

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Numéro attribué par la bibliothèque

THÈSE

pour obtenir le grade de

DOCTEUR DE L'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Spécialité : Imagerie, Vision et Robotique
Ecole Doctorale : Mathématiques, Sciences et Technologie de l'Information

présentée et soutenue publiquement
par

Thi Thanh Hai TRAN

le 31 Mars 2006

Titre :

ETUDE DE LIGNES D'INTERET NATURELLES POUR LA REPRESENTATION D'OBJETS EN VISION PAR ORDINATEUR

JURY

M. James L. Crowley	Président
M. Michel Dhome	Rapporteur
M. Justus Piater	Rapporteur
M. Augustin Lux	Directeur
Mlle. Hoang Lan Nguyen Thi	Co-directrice
M. Eric Marchand	Examinateur

Thèse co-tutelle préparée dans le laboratoire GRAVIR – IMAG au sein du projet PRIMA
et le Centre de Recherche International MICA – Hanoi VietNam

Remerciements

Tout d'abord, je tiens à remercier mon directeur de thèse, M. Augustin Lux qui m'a introduit dans ce domaine de recherche depuis mon DEA. Je le remercie pour m'avoir aidé, encouragé et renforcé les motivations qui m'ont manqués. Je voudrais le remercie également pour sa compréhension, sa gentillesse et son soutien tout au long de cette thèse. Je voudrais lui exprimer tous mes respects et mes reconnaissances pour m'avoir appris le bon esprit de la recherche.

Je tiens à remercier Mlle Hoang Lan Nguyen Thi, ma co-directrice, pour les discussions utiles sur mon travail de thèse ainsi que ses conseils sur la rédaction du manuscrit. Je la remercie encore pour son soutien et son encouragement persistant depuis j'étais une étudiante de l'IPH.

Je tiens à remercier les personnes qui me font honneur de participer à mon jury : M. James L. Crowley comme président, M. Michel Dhome et M. Justus Piater comme rapporteurs qui on contribué à évaluer mes travaux de thèse, M. Eric Marchand comme examinateur.

Je tiens à remercier M. James L. Crowley et M. Augustin Lux pour m'avoir financié pendant la thèse et m'a donné des matériels essentiels pour réaliser cette thèse.

Je remercie mes collègues qui m'ont aidé à relire et clarifier mon manuscrit : Olivier Bertrand, Nicolas Gourier, Sofia Zaidenberg, Alain Boucher, Daniela Hall. Plus particulièrement Olivier Bertrand pour les discussions utiles afin d'améliorer le contenu du manuscrit et Alain Boucher pour avoir passé son temps aux longues discussions qui m'ont motivées à l'application de la détection de texte.

Je tiens à remercier tous les membres de l'équipe PRIMA et ceux de MICA pour une très bonne ambiance de travail et les amies qui rendent la vie moins dure.

La thèse est une période difficile de la vie mais plein de bonheur, surtout quand on arrive à la fin. Je voudrais remercier mes parents, qui ont toujours une forte confiance de ma réussite. Et de la profondeur du coeur, je voudrais exprimer tous mes remerciements à Dang Minh Dung pour m'avoir partagé la vie et lui donné le sens.

Résumé

Extraction de caractéristiques est une étape essentielle dans tous les systèmes visuels. Depuis quelques années, les recherches se focalisent sur un type de caractéristique visuelle appelé “point d’intérêt” avec grande richesse de résultats. Cependant, les points d’intérêt se prêtent mal à une modélisation structurelle. Cette thèse consiste à étudier un type de caractéristiques qui permet de représenter la topologie des structures dans l’images : les crêtes. Les points de crêtes sont les extrema directionels sur les surfaces des images lissées dans l’espace d’échelle. Ils sont détectés par un opérateur de Laplacien. Ces points discrets sont étiquetés pour former des lignes de crêtes à l’aide d’un algorithme d’analyse de composantes connexes. Les lignes de crêtes obtenues caractérisent l’axe central des structures ainsi que leurs tailles.

Les crêtes, par leur nature “ligne”, sont très utiles pour représenter les objets structurés comme par exemple la silhouette de l’être humain ou la ligne de texte. Nous montrons l’utilisation de crête dans deux applications : la modélisation de personnes et la détection de textes. Chaque silhouette de l’être humain est représenté par des crêtes significatives correspondant aux torse et jambes. Cette représentation aide à caractériser la configuration de la personne en mouvement à chaque instant donné. Chaque texte est modélisé par une crête longue à une échelle grande et nombreuses crêtes courtes non-parallélisme à une petite échelle. La modélisation structurelle du texte est générique pour plusieurs types de texte et indépendant de l’orientation du texte.

Mots-clés

Espace d’échelle, Crête, Représentation structurelle, reconnaissance d’objet.

Natural line extraction for Object Representation in Computer Vision

Abstract

Feature extraction is a crucial step in all visual systems. Since a few years, on aims to look for features like interest points, key points or salient points. Local feature point based methods have shown to be very efficient for object recognition. However, points in a high-dimensional feature space do not allow abstract representation of object shape, and, even if they are invariant to imaging conditions, they are not good at generalization. The work represented in this dissertation concerns a very precise question : Which roles can a ridge play for object representation ? We study ridges on surfaces associated to smoothed images in scale-space. Ridge points are directional local extrema, detected by using Laplacian operator. These points are labelled to build ridges lines, useful for object representation.

Ridges, by theirs “line” nature, are very useful to represent structural object line human silhouette or text line. We proposed to modelize human and text using some significant ridges. A human is represented by ridges corresponding to torso and legs of the human. This representation permits to analyze human movement via his configuration. A text line is modelized by a long ridge at coarse scale corresponding to the center line of the text line and several smaller ridges at finer scale corresponding to character skeletons. Ridges at small scale must not be parallel to the main ridge. This text model is generic for all types of text and independent with text orientation.

Keywords

Scale-space, Ridge, Structural representation, Object recognition.

Table des matières

Introduction

Pour construire des systèmes intelligents, il est nécessaire de leur fournir des capacités de perception. Parmi ces capacités la vision est la plus puissante pour capturer l'information sur l'environnement. Si nous dotons les machines de la faculté de voir, leurs décisions et réactions vont gagner en effectivité et efficacité.

La capacité de "voir" ne se limite pas à la capture des images, mais nécessite aussi l'analyse, inclut des raisonnements, afin de comprendre pour pouvoir prendre des décisions. Avec le developpement rapide des technologies, l'acquisition d'images est devenue une tâche banale. Leur compréhension et interprétation reste en revanche toujours un défi dans le domaine de la vision par ordinateur.

La principale difficulté qu'un système de vision doit surmonter est de s'affranchir de la représentation d'une image sous forme de matrice de pixels. Quel type d'information doit y être détectée sans ou avec très peu de connaissances de l'environnement ? Comment l'obtenir et la représenter pour permettre de comprendre la scène ?

Le premier système de vision artificiel, mis au point dans les années 60, fonctionne dans un monde lui aussi très artificiel, le "monde des blocs" (appelé en anglais "blocksworld"). La donnée d'entrée de ce système est l'image d'une scène qui ne contient que des objets polyédriques. L'analyse topologique de l'image, en détectant des segments de droite et en les reliant via des sommets permet de l'apparier avec le modèle 3D des objets et de redessiner la scène à partir d'un point de vue aléatoire.

50 ans plus tard, on obtient aujourd'hui des résultats impressionnant en vision par ordinateur. On a réussit à construire des systèmes de vision artificielle qui fonctionnent dans un environnement réel, même très complexe et dynamique. On peut citer par exemple la navigation autonome (eg. Mars rover[1]), l'asservissement visuel[PM04, HAMMC05, CMPC06], le suivi de personnes[BD01, RFZ05, ZN04, CHRC04], la recherche d'images d'une grande base[MN95, SM97, PCpS99, RLSP05], la modélisation de personnes[LY95, Bau95, HHD98a, ZNL01], etc.

Cette réussite est due à de grandes avancées concernant la recherche de caractéristiques à la fois significatives et discriminantes dans une image. Significatives pour permettre une représentation d'objets compacte et informative, discriminantes pour pouvoir distinguer des objets différents. Ces caractéristiques, qu'il s'agisse de régions, de lignes de contour ou de points d'intérêt, sont extraites par des techniques de segmentation, de détection de contraste ou de changements significatif du signal.

Si les caractéristiques telles que les régions, les lignes de contours ou les arrêtes ont considérablement contribué dans les années 80 à la modélisation d'objets à partir d'images, les points d'intérêt occupent le premier plan de la recherche de ces dix dernières années. Décrivant des changements importants du signal 2D de façon locale, ils sont souvent étudiés à travers l'espace-échelle ce qui permet d'obtenir

[1] http ://www.space.com/marsrover/

l'invariance aux changements d'échelle. Par ailleurs, ils peuvent être relocalisés avec plus de précision en espace et en échelle, via des techniques de dérivées.

Les points d'intérêt cumulent les propriétés que l'on souhaite d'une caractéristique : une bonne localisation, l'invariance aux changements d'échelle et de luminosité. Les résultats de reconnaissance ou de matching utilisant des descripteurs basés sur les points d'intérêt sont impressionnants. Les points d'intérêt ont révolutionné la vision par ordinateur.

Dans le but de chercher des caractéristiques pour aider à comprendre la scène, il est évident qu'aucune caractéristique citée ci-dessus peut représenter tout type de scènes. Si les contours, ou les régions sont fortement reliés à des structures physiques dans la scène, les points d'intérêt sont loin de permettre une représentation sémantique de l'objet. Leur nature ponctuelle donne de la souplesse aux différentes techniques de description, mais l'information qu'ils contiennent reste encore très explicite. Les points d'intérêt se prêtent mal à la modélisation des formes d'objets.

Dans la recherche de caractéristiques, une question que l'on se pose est que la suivante : Depuis la naissance de la géométrie, les points et lignes ont été étudiés comme deux primitives de base. Tandis que les points d'intérêt ont convaincu le monde de la vision artificielle au cours de cette dernière décennie, que peuvent apporter leur duaux : les lignes d'intérêt ?

La figure 1 montre trois exemples de points d'intérêt et de lignes d'intérêt sur les mêmes images. Les points d'intérêt représentent des endroits texturés dans l'image tandis que les lignes d'intérêt représentent la forme des structures allongées dans l'image (e.g. cardes d'un vélo, routes dans l'image aérienne, branches d'un arbre). Décrire un vélo, une route ou un arbre, par des lignes naturelles, sera sémantiquement beaucoup plus intéressant que par des points.

Dans une scène, la plupart des composantes constituant des objets sont de forme allongée. Dans le cas des composantes arrondies, nous pouvons les considérer comme des structures allongées soumises à une compression en un blob. De cette manière, les lignes d'intérêt sont génériques pour la représentation de toute forme de structure, ce qui est bien pour la reconnaissance générique où on a besoin d'une caractérisation plus abstraite de la forme. Intéressés par cette propriété, notre travail de thèse consiste à exploiter les rôles que peuvent jouer les lignes d'intérêt pour la représentation d'objets et leur reconnaissance.

La reconnaissance d'objets est un thème fondamental qui se présente sous les formes les plus diverses dans toutes les applications de la vision artificielle. Quelque soit la structure d'un système de reconnaissance, il doit confronter des problèmes fondamentaux : il faut représenter les objets sous forme de modèles formels et informatiques, ce qui pose d'abord le problème de la nature des caractéristiques visuelles à utiliser, le problème de la structure d'un modèle et d'une base de modèles ensuite, et enfin le problème de l'apprentissage de ces modèles.

Il faut aussi reconnaître à partir d'une image : cela pose le problème de traitement d'image pour recouvrer les caractéristiques visuelles, le problème de l'appariement avec les modèles, le problème du raisonnement pour arriver à une interprétation de l'image.

Une ligne d'intérêt étant une entité sensiblement plus complexe qu'un point, il n'est pas question d'aborder pour les lignes d'intérêt tous les thèmes de recherche cités ci-dessus ; nous avons cherché des applications plus simples, et plus directement exploitables, que ne l'est, par exemple, la recherche des routes dans une image aérienne : notre travail est une première exploration, allant de la définition mathématique et algorithmique des lignes d'intérêt jusqu'à l'étude de deux applications : la recherche de formes "être humain", et la détection d'écritures.

Dans ces conditions, les contributions principales de cette thèse sont :

FIG. 1 – Point d'intérêt vs. ligne d'intérêt. Première colonne : les 100 points d'intérêt les plus forts en énergie fournis par le détecteur Harris-Laplacien. Deuxième colonne : lignes d'intérêt (en bleu et en rouge) détectées à échelle fixe (première et deuxième ligne) et à l'échelle caractéristique (dernière ligne).

– Une méthode de détection multi-échelle de lignes d'intérêt naturelles, utilisant le Laplacien du Gaussien et la matrice Hessienne.
– Une approche structurelle à la recherche de forme "être-humain", dans laquelle les lignes d'intérêt naturelles correspondent au corps et aux membres d'une personne.
– Une nouvelle approche de modélisation d'écritures, dans laquelle les lignes d'intérêt naturelles, à différentes échelles, modélisent à la fois la ligne et les trait d'écriture.

Organisation du rapport

Le contenu du rapport s'organise de la manière suivante :

Le chapitre 1 présente un état de l'art partiel de la reconnaissance générique. Nous présentons d'abord les caractéristiques visuelles les plus utilisées. Puis, nous exposons les méthodes de reconnaissance classiques.

Le chapitre 2 rappelle d'abord quelques bases de la géométrie différentielle des surfaces et la représentation multi-échelle de l'image. Nous présentons plusieurs définitions et plusieurs méthodes de détection des points sur lignes d'intérêt naturelles, puis nous proposons notre méthode de détection basée sur l'opérateur Laplacien du Gaussien. Cet opérateur connu pour la détection de contours et de pics est aussi efficace pour les lignes d'intérêt naturelles.

Le chapitre 3 décrit une méthode pour identifier une région contenant des personnes par des lignes d'intérêt naturelles représentant torse et jambes. Les modèles de personnes sont appris par un algorithme de type K-Means puis la classification s'effectue en comparant le modèle construit de nouvelle imagette avec les configurations apprises. Une comparaison avec deux autres méthodes implémentées dans notre équipe est également présentée : une méthode à base d'histogramme de gradients et une méthode utilisant des mémoires linéaires auto-associatives. Cette étude a été réalisé dans le contexte du projet CAVIAR[2].

Le chapitre 4 étudie l'utilisation des lignes d'intérêt naturelles pour modéliser une région de texte. Les méthodes de détection de texte dans la littérature sont d'abord présentées. Ces méthodes se basent principalement sur les caractéristiques de couleur, de texture, de contour. Nous proposons une nouvelle approche qui modélise l'organisation structurelle des lignes et des caractères à travers plusieurs échelles. Une région de texte est caractérisée par des lignes de crête à une grande échelle - les lignes du texte - et des lignes d'intérêt naturelles courtes à des échelles plus petites (squelettes des caractères). Les problèmes de choix des échelles et de réglage des seuils sont abordés.

Le chapitre 5 présente un travail plus prospectif et plus ambitieux : l'utilisation des lignes d'intérêt naturelles pour la représentation hiérarchique structurelle d'objets complexes. Nous présentons la représentation proposée et les premiers résultats expérimentaux de reconnaissance, par un algorithme de mise en correspondance de graphes hiérarchiques.

Le chapitre 6 donne la conclusion sur les travaux réalisés et quelques perspectives de l'utilisation de lignes d'intérêt , accompagnant par des descripteurs numériques pour une modélisation hiérachique plus efficace de l'objet.

[2] http ://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm

Chapitre 1

Modélisation par caractéristiques visuelles

Ce chapitre présente un état de l'art des méthodes de modélisation d'image utilisant des caractéristiques visuelles pour la reconnaissance des classes d'objets en vision par ordinateur. Dans ce sujet vaste nous nous intéressons particulièrement à quatre questions :

1. La nature des caractéristiques visuelles[1] employées.
2. La représentation des modèles d'objet.
3. Les techniques de construction des modèles, notamment les techniques d'apprentissage.
4. la mise en correspondance modèle-image, qui est le problème algorithmique à proprement parler de la reconnaissance.

Ces quatre questions sont étroitement liées. Par exemple : dans le cas où les pixels servent directement de caractéristiques, le modèle est représenté par une imagette, et la mise en correspondance modèle-image se fait par corrélation. Autre exemple : si l'on représente un modèle par un histogramme, la mise en correspondance est un calcul de distance entre histogrammes.

Dans toute cette thèse, nous plaçons les caractéristiques visuelles au centre de notre intérêt : le choix des caractéristiques conditionne la structure des modèles, et par là leur construction ; il également détermine les algorithmes de reconnaissance, en ce sens qu'il définit les types des paramètres d'entrées et du résultat.

Ce chapitre commence par une présentation des caractéristiques visuelles les plus utilisées, puis décrit quelques méthodes de représentation fondées sur ces caractéristiques. Cet état de l'art motive la suite de notre travail : pour améliorer l'expressivité et la robustesse des caractéristiques classiques, et pour pallier la faiblesse des techniques numériques à la catégorisation abstraite des formes nous allons étudier un type de caractéristique peu employée actuellement, les lignes d'intérêt naturelles, et développer une méthode de représentation hiérarchique structurelle pour les modèles d'objets.

1.1 Caractéristiques visuelles

Un système de reconnaissance visuelle peut, à la limite, s'affranchir du calcul de caractéristiques, et directement utiliser les pixels de l'image pour représenter les objets, par exemple les méthodes de

[1] notre traduction du terme anglais *features*

reconnaissance par apparence proposées par Swain et Ballard [Swa91], Murase et Nayar [MN95], Ohba et Ikeuchi [OI96, OI97], ou encore Schiele[SC96c, SC00, LS03]. Dans ces systèmes, la représentation d'objets utilise l'image entière ou un ensemble d'imagettes. Cette représentation est simple ; en contrepartie, elle est redondante, elle nécessite une grande quantité de mémoire pour le stockage, et le temps d'appariement augmente linéairement avec le nombre de modèles. Par ailleurs, son expressivité restera limitée.

Très généralement, l'extraction de caractéristiques d'une image a pour but d'extraire des informations essentielles, significatives, et discriminantes pour une représentation compacte et pour une reconnaissance efficace et rapide. En plus, les caractéristiques devraient correspondre à une structure de l'objet ou un évènement physique de la scène, permettant une représentation plus "sémantique" de l'image.

La performance d'un système de reconnaissance d'objets dépend fortement des caractéristiques utilisées. Une caractéristique peut être efficace pour un type d'application mais inefficace pour une autre. Pour juger une caractéristique, il faut prendre en compte des critères très différents, et notamment *l'expressivité, la répétabilité, la robustesse, l'efficacité*.

1. *Expressivité* : L'expressivité d'une caractéristique est son pouvoir de modélisation. Par exemple, les segments de droite représentent bien des objets manufacturés tandis que les pics du Laplacien[2] sont une bonne représentation pour les structures rondes. Une caractéristique est générique si elle permet de représenter plusieurs types de structures. Ce pouvoir de modélisation se juge aussi par la capacité de discriminer entre différents objets.

2. *Répétabilité* : La répétabilité est définie comme le fait qu'un même point physique d'un objet visible dans deux images soit détecté dans deux images [Sch92, dV99, HLS02, MS05]. La répétabilité est l'invariance aux transformations telles que la rotation et la translation, aux changements d'éclairage et d'échelle. Cette propriété est manifestement cruciale pour une représentation de modèles d'objet. Comme l'invariance parfaite ne peut rarement être atteinte, on introduit des formes dégradées, comme la ϵ*-répétabilité*.

3. *Robustesse* : Une caractéristique est robuste si elle est faiblement influencée par le bruit. On appelle bruit ici tout ce qui perturbe la scène capturée dans l'image, pour une cause matérielle ou logicielle, par une source "interne" ou "externe". La conséquence du bruit est une mauvaise qualité de l'image : structures floues ou cassées, donc difficiles à reconnaître même par l'oeil humain.

4. *Efficacité* : Le calcul d'un type de caractéristique dans une image doit se faire avec des algorithmes efficaces - en "temps réel". Ce critère dépend en partie des progrès de la technologie, mais il exclut, par exemple, des algorithmes itératifs au niveau des pixels, d'isomorphisme de graphes, etc.

Une caractéristique peut être un pixel (ie. les points d'intérêt), une ligne (les lignes de contours, de squelettes ou de crête) ou une région dans l'image. Le bon cadre de détection de caractéristiques n'est pas la seule image de pixels, donnée originale, mais l'image augmentée de son *espace d'échelle*. Il s'agit de détection de caractéristiques à multi-échelles ou à l'échelle caractéristique. Avant d'étudier des caractéristiques, nous rappelons d'abord la notion de l'échelle et l'échelle caractéristique. La représentation de l'image dans l'espace d'échelle sera présentée en détail dans le chapitre 2.

[2] blobs en Anglais

1.1.1 Echelle

Un objet dans le monde réel est typiquement modélisé par des composantes de tailles très différentes, et dont les images correspondent à des bandes de fréquences spatiales différentes. Par exemple dans une image d'un arbre, le tronc peut avoir la taille de 50 pixels en largeur tandis que les branches n'auront que 10 pixels (voir la figure 1.1). La description de l'objet par des composantes primitives doit assurer la correspondance des caractéristiques détectées avec des composantes physiques de l'objet. Pour cela, chaque caractéristique doit être détectée à une échelle appropriée à sa taille.

FIG. 1.1 – Exemple d'un objet composé des structures de tailles différentes.

La détection de caractéristiques d'une taille quelconque est réalisée par des détecteurs paramétrables par l'échelle. Pour cela, un détecteur doit être une fonction de la position et de l'échelle $f(x, y; \sigma)$. Etant donné une caractéristique dans une image, l'échelle pour laquelle un détecteur donne une réponse maximale est appelée *échelle caractéristique* (voir la figure 1.2).

Dans [Lin94], Lindeberg a proposé une méthode de sélection automatique pour l'échelle caractéristique des pics et des crêtes en cherchant les maxima en énergie des mesures de caractéristiques correspondantes. Plus spécifiquement, l'échelle caractéristique d'un pic est l'échelle à laquelle la magnitude du Laplacien est maximale. L'échelle caractéristique d'un point de crête est l'échelle à laquelle la plus grande courbure principale de la surface locale associée est maximale dans la direction principale correspondante.

1.1.2 Région

La caractérisation d'une image par un ensemble de régions correspond à la supposition que les objets réels sont constitués d'un certain nombre de composantes distinctes. Par exemple un vélo est composé de 2 roues, un cadre, un guidon, etc. Chaque composante est une région connexe des points ayant des propriétés particulières.

Dans la littérature sur les méthodes de reconnaissance générique, il y a deux approches à la détection et à l'utilisation des régions : la première approche se base sur l'observation qu'une composante de l'objet

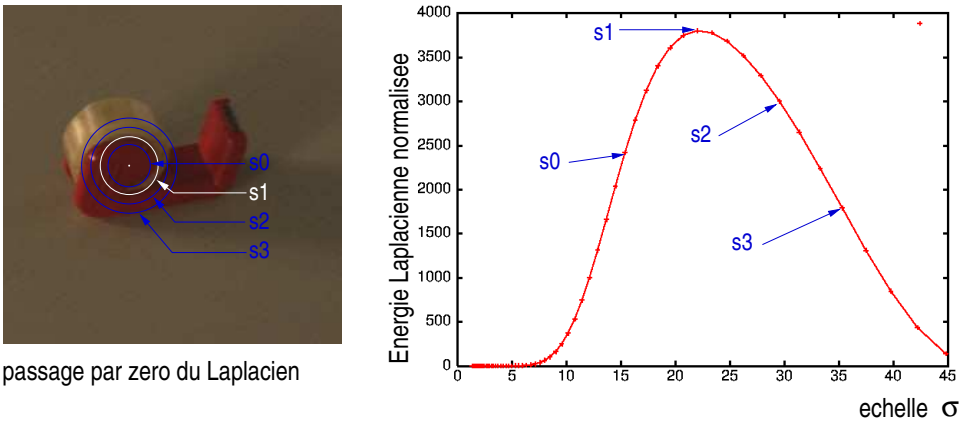


FIG. 1.2 – L'image d'un rouleau de scotch et le profil de l'énergie du Laplacien normalisée à l'échelle. Quand l'échelle est égale à la taille de l'objet s_1, l'énergie de Laplacien est la plus grande [Hal01]

est une région dont les pixels sont homogènes selon un certain critère (intensité, couleur, texture [RA01, KD05]) ; la deuxième approche se base sur une sélection des positions intéressantes dans l'image puis considère les régions autour des points d'intérêt [AC99, HdVC00, HCCdV00, Low04, FPZ04, DS05].

La première approche utilise les techniques de segmentation d'image classiques : seuillage de l'intensité (utilisant l'histogramme) ou des techniques plus sophistiquées comme la segmentation "split-merge" [CP80], la croissance de régions [Zuc76] etc. (pour un résumé voir [FM81, HS85, PP93]).

D'une façon générale, ces techniques n'abordent pas la notion d'échelle. Une région détectée correspond souvent à une composante physique de l'objet, mail peut parfois résulter d'une fusion de plusieurs composantes de différents objets (voir la figure 1.3). Un inconvénient majeur des ces approches classiques est leur sensibilité au bruit, due à l'emploi de seuils. Cela a provoqué des recherches sur le paramétrage automatique de seuils.

La figure 1.3 montre le résultat de segmentation d'image par la technique basée sur le champs de vecteurs de contour développée par Sumengen[3] [Sum05]. Clairement, les régions significatives représentent des structures de l'objet et de la scène. Cependant, la séparation n'est pas complètement satisfaisante car quelques composantes du vélo sont fusionnées avec celles du fond.

La deuxième approche se base sur les techniques de détection des points d'intérêt présentées ci-dessous. En considérant que les structures intéressantes se trouvent autour des points d'intérêt dans l'image, la détection de points d'intérêt puis la détermination d'imagettes les incluant fournissent un ensemble de régions intéressantes pour caractériser l'objet. Parfois, ces régions peuvent ne couvrent pas tout l'objet. Mais tel est le but de l'extraction de caractéristiques : rendre compacte la représentation de l'objet en gardant seulement quelques composantes informatives.

[3]Le programme se trouve à http ://aakash.ece.ucsb.edu/imdiffuse/segment.aspx

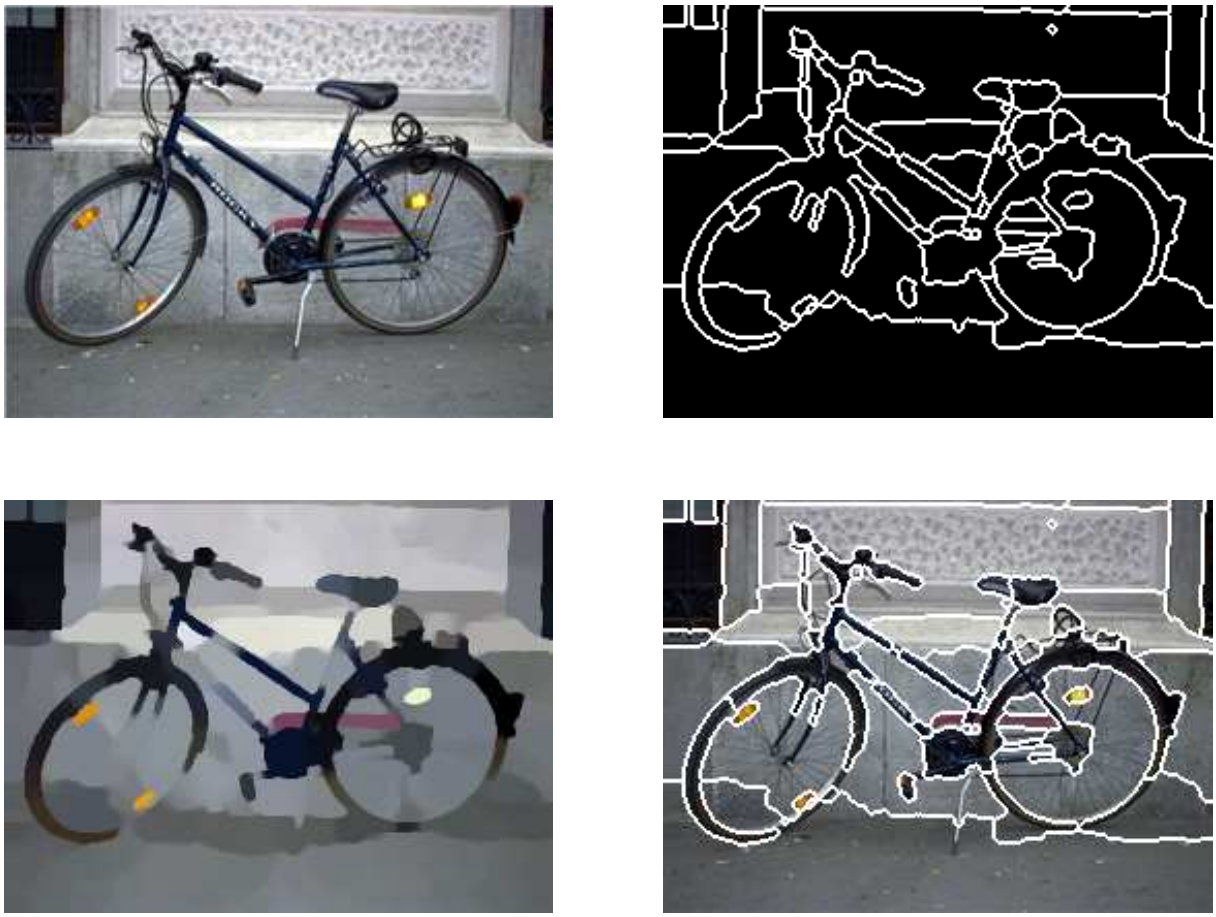


FIG. 1.3 – De gauche à droite, de haut en bas : Image d'un vélo. Plan de contour. Image diffusée. Image segmentée à l'échelle $\sigma = 3$ selon l'algorithme proposé par Sumengen [Sum05].

1.1.3 Point significatif

Il n'est pas exagéré de dire que le concept de point d'intérêt, ou point significatif, a révolutionné la vision par ordinateur, en introduisant des caractéristiques visuelles certes moins intuitives que les régions ou les segments de droites abondamment utilisées auparavant, mais remarquablement expressives pour une grande variété d'objets. L'idée de base a donné lieu à de nombreux travaux, avec des terminologies souvent différentes. D'une façon générale, nous subsumons sous le terme "point significatif" plusieurs types de points tels que le coin, la jonction [Sch92], le point saillant [kB01], le point clé [Low04], etc. Dans tous ces cas, il s'agit des points où le signal image a des propriétés très particulières, comme un changement brusque de l'énergie dans toutes les directions, ou un maximum d'entropie, etc. Nous présentons ci-dessous quelques types de point significatif les plus étudiés et utilisés actuellement dans le domaine de reconnaissance générique.

Point d'intérêt

Les points d'intérêt sont des points où le signal de l'image accuse un changement brusque dans les deux dimensions spatiales. Des exemples sont les coins, les jonctions en T, et aussi les endroits où la texture varie fortement. Pour détecter le changement bidimensionnel, plusieurs approches ont étés proposées, depuis des années 80.

Selon une évaluation quantitative récente des détecteurs de points d'intérêt réalisée par Mikolajczyk *et al.* [MS04b, MS04a, MS05], le détecteur Harris-Laplacien adapté à l'échelle serait le meilleur détecteur de point d'intérêt selon les critères de répétabilités et de discriminalité face au bruit, aux rotations, translations, changements d'échelle et de la lumière. Un point d'intérêt détecté par l'opérateur Harris-Laplacien est un point qui satisfait deux critères :

$$det(C) - \alpha trace^2(C) > seuil_h \text{ et}$$

$$L(x,y;\sigma_{n-1}) < L(x,y;\sigma_n) > L(x,y;\sigma_{n+1}) \cap L(x,y;\sigma_n) > seuil_l$$

où $seuil_l$ est un seuil de l'énergie de Laplacien et α, $seuil_h$ sont les seuils expérimentalement choisis. C est une matrice proposée par Harris adaptée à l'échelle[4] :

$$C(x,y;\sigma) = \sigma^2 \begin{bmatrix} {L_x}^2(x,y;\sigma) & L_x(x,y;\sigma)L_y(x,y;\sigma) \\ L_x(x,y;\sigma)L_y(x,y;\sigma) & {L_y}^2(x,y;\sigma) \end{bmatrix}$$

où σ est l'échelle de détection. La figure 1.4 montre les points détectés par l'opérateur d'Harris-Laplacien superposés sur l'image du vélo de la figure 1.3. Les points détectés se trouvent aux endroits correspondant à des structures importantes du vélo : le cadre, les roues, le guidon. Ils sont les points sur lesquelles nous devons nous concentrer pour la reconnaissance.

Notons cependant, que ces points sont déconnectés les uns des autre, ce qui complique une interprétation sémantique.

Point clé

A côté des points d'intérêt, les points clés, introduits par Lowe 2004 ? ? [Low04] ont été largement utilisés pour la classification des images [Low04, MHSTS05, OFPA04a]. Dans [Low04], un point clé est un maximum dans l'espace d'échelle. Plus précisément, un point $(x,y;\sigma)$ est identifié comme point clé à l'échelle σ si l'énergie de différence des Gaussiens en ce point est plus grande que celle en 26 points voisins dans l'espace d'échelle (figure 1.5). Comme la différence de Gaussiens est une approximation du Laplacien : $G(x,y;k\sigma) - G(x,y;\sigma) \approx (k-1)\sigma^2\nabla^2 G$, les points clés sont en fait un sous-ensemble des pics - des maxima dans 4 directions spatiales et dans la dimension d'échelle - utilisés par Lindeberg [Lin94].

Un point clé est caractérisé par la direction et la magnitude du gradient. La figure 1.6 montre les points clés détectés sur l'image du vélo. Les flèches blanches représentent le vecteur du gradient aux points clés [5]. On constate que les points clés sont détectés sur les contours, et aux centres de régions uniformes.

[4] pour les notations, voir le chapitre 3.

[5] Le programme de détection est fourni par Lowe : http ://www.cs.ubc.ca/spider/lowe/home.html

FIG. 1.4 – Les 100 points les plus forts en énergie de Laplacien détectés par le détecteur Harris-Laplacien.

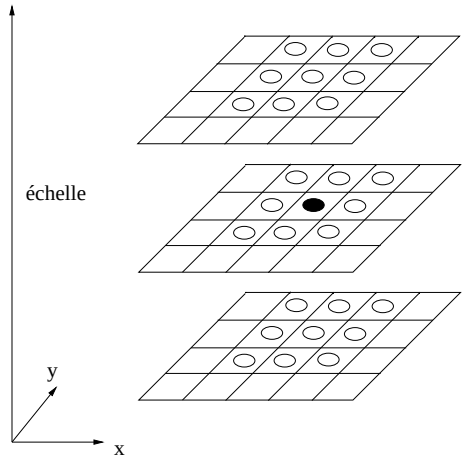


FIG. 1.5 – Point d'intérêt selon le critère proposé par Lowe [Low04].

FIG. 1.6 – Détection de points clés sur l'image du vélo.

Point saillant

Un type de point introduit par Gilles en 1998 [Gil98] puis développé par Kadir et Brady [kB01] dans l'espace d'échelle est appelé *point saillant* dans l'image. Il s'agit des points *uniques* de l'objet pour lesquels la discriminance est maximisée.

La "discriminance" d'un point est une mesure d'entropie locale, calculée dans une région autour du point. L'algorithme de détection de point saillant dans l'espace d'échelle proposé par Kadir et Brady est le suivant :

Pour chaque point (x,y) :

- à l'échelle s dans l'intervalle $[s_{min}, s_{max}]$
  - Calcul des descripteurs locaux dans le voisinage R_{xy} de (x, y) $D = \{d_1, d_2, ..., d_r\}$ où d_i est l'intensité en point i dans le voisinage R_{xy}.
  - Estimation de la fonction de densité de probabilité des descripteurs : $P_{D,R_{xy}}(d_i)$. La probabilité est calculée en utilisant l'histogramme de l'intensité.
  - Calcul de l'entropie locale : $H_{D,R_{xy}} = -\sum P_{D,R_{xy}}(d_i) log_2 P_{D,R_{xy}}(d_i)$.
- Sélection des échelles S auxquelles l'entropie est locale maximale.
- Pondération des valeurs de l'entropie à des échelles sélectionnées par la somme de la différence absolue des probabilités de descripteur local autour de S.

Kadir et Brady ont montré la signification des points saillants détectés dans les images expérimentales. Ces points correspondent bien à des objets d'intérêt dans l'image tels que les personnes, les voitures etc. Le détecteur des points saillants est utilisé dans les travaux de Fergus et al. [FPZ04] et Dorko et Schmid [DS05] pour la détection des composantes significatives de l'objet dans l'image et la reconnaissance des classes d'objets.

1.1.4 Contour

Le contour est une caractéristique qui décrit les structures importantes dans l'image et établit des indices afin d'obtenir des structures de la scène originale. Intuitivement, un contour est *une ligne* du long de laquelle se produit une brusque variation d'intensité de lumière dans une direction. Ainsi, le contour est une structure uni-dimensionnelle. Il correspond à la discontinuité d'ordre 1 de la surface associée à l'image originale.

Un contour dans une image correspond à un bord d'un objet physique, ou à une structure quelconque (ie. ombre). Les contours produits par les effets de lumière ne sont pas intéressants pour la représentation de l'objet mais restent significatifs pour l'interprétation de l'image.

Comme le contour correspond à une discontinuité du signal, on utilise des opérateur de dérivée pour leur détection. On peut trouver de très nombreux algorithmes de détection de contours, basés sur la recherche des maxima de la dérivée première ou des passages par zéro de la dérivée seconde [Mar82, Can83]. Dans cette thèse, nous utilisons les passages par zéro du Laplacien du Gaussien $\nabla^2 G(x, y; \sigma)$ proposé par Marr [Mar82]. Cet opérateur est paramétré par l'échelle σ. Ainsi, les contours peuvent être détectés à différentes échelles.

FIG. 1.7 – Détection de contour à l'échelle $\sigma = 4$. A gauche : image de points de contour. A droite : superposition des points de contour sur l'image originale.

Les figures 1.7, 1.8 montrent le résultat de cet opérateur à plusieurs échelles. On constate que les contours correspondent aux bords des objets dans la scène, et aux effets de lumière qui provoquent le changement d'intensité (ombre). La détection à plusieurs échelles fournit les contours des petites structures ainsi que ceux de structures globales de la scène. Pourtant, la combinaison de ces informations en une description unique est difficile à cause de comportements difficilement interprétables des contours à travers les échelles ("primal sketch" de Marr).

1.1.5 Squelette, Crête

Pour une représentation structurelle de l'objet, les caractéristiques décrivant la topologie de l'objet sont préférables. A part le contour, le squelette et la crête sont les caractéristiques intrinsèques de l'objet.

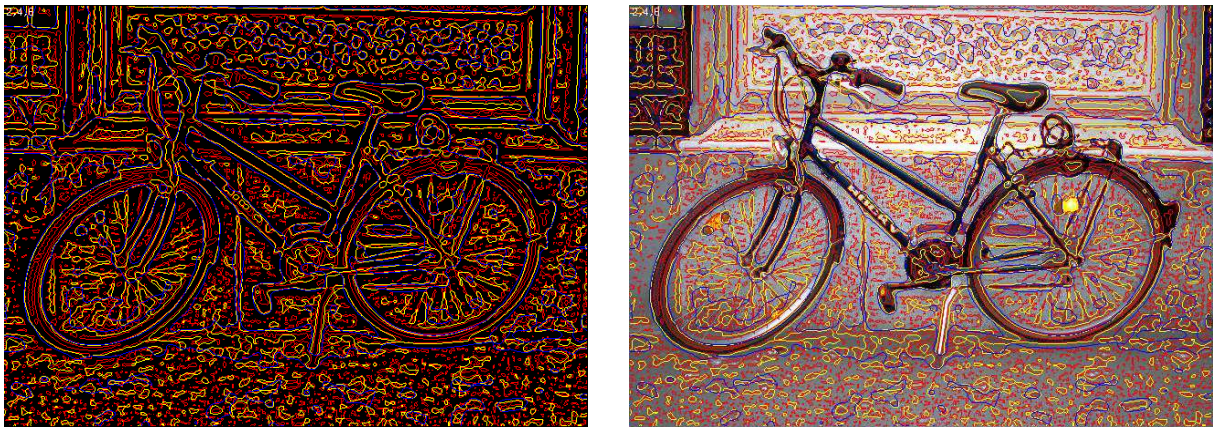


FIG. 1.8 – Détection de contours à 3 échelles $\sigma = 2, 4, 8$. Les couleurs représentent l'échelle de détection : rouge $\sigma = 2$, vert $\sigma = 4$, bleu $\sigma = 8$.

La notion de squelette est introduite par H. Blum [Blu67] comme le résultat d'une transformation de l'axe médian[6]. La transformation de l'axe médian détermine les points de contour les plus proche de chaque point dans une région S donnée. Un point à l'intérieur de S appartient au squelette si au moins 2 de ses points les plus proches sont des points de contour. Il existe une définition formelle de squelette basée sur la notion de boule maximale. Le squelette d'une forme S est l'ensemble des centres des boules maximales dans S.

Le squelette a plusieurs propriétés intéressantes. Il est théoriquement invariant par transformation linéaire (translation, rotation, changement d'échelle). La squeletisation[7] conserve les propriétés topologiques de la forme d'origine ainsi que ses propriétés géométriques. Pourtant, une propriété générale considéré comme défaut est que la squeletisation est une transformation semi-continue. En effet, la moindre perturbation dans le contour ou au sein de la forme peut produire la création d'une branche importante dans le squelette. La squelette est ainsi très sensible au bruit.

Le squelette connaît plusieurs applications telles que la reconnaissance des formes, la modélisation de solides pour la conception et la manipulation de formes, l'organisation de nuages de points, la recherche de chemins, les animations, etc. Elle est utilisée en médecine et en biologie depuis sa création, ainsi qu'en minéralogie. Des applications ont été trouvées dans l'indexation d'images dans les bases de données et en compression. Il existe également quelques applications en architecture et en urbanisme, dans le cadre d'analyse morphologique.

Malgré sa large utilisation en plusieurs domaines cité ci-dessus, le squelette n'est pas une bonne caractéristique pour la reconnaissance d'objets. La raison en est que la squeletisation nécessite que l'image soit binaire. La binairisation de l'image est en fait un processus de segmentation d'image, qui reste un problème primitif. En outre, la définition de squelette via la squelitisation ne permet pas toujours de trouver un squelette et la création de branches inattendues à cause d'une petite perturbation empêche une représentation fiable de l'objet.

[6] Medial Axis Transformation (MAT)

[7] Un ensemble d'algorithmes utilisées en analyse de forme qui consistent à réduire une forme en un ensemble de courbes

FIG. 1.9 – Les crêtes négatives détectées de l'image d'un vélo à l'échelle $\sigma = 4\sqrt{2}$. Nous constatons la signification de caractéristiques type crête pour la représentation des structures de l'objet par rapport à des régions, contours ou points d'intérêt.

Récemment, plusieurs travaux se sont concentrés sur la recherche des caractéristiques "crêtes"[8] [SV52, EGM+94, Ebe94, Ebe96] grâce à leur multiples avantages : D'abord, la crête est définie sur une surface mathématique. Elle peut être la surface de l'image originale ou de l'image lissée. La détection de crête n'a pas besoin donc de segmentation. Ensuite, la crête 2D conserve la topologie de la forme des structures. Enfin, l'étude de la crête dans l'espace d'échelle permet une présentation grossière - détaillée de l'objet. Par rapport aux squelettes, les crêtes rapportent plus d'information de la forme et permettent plusieurs algorithmes efficaces pour la détection ainsi que la flexibilité dans la représentation et la reconnaissance.

La crête a été utilisée pour l'analyse d'image médicale, plus précisément la mise en correspondance [EGM+94, Ebe94, Ebe96, SAN+04, SKA+02], la caractérisation de structures [LL00] ou la représentation de l'objet [Cro81]. Nous proposons dans cette thèse d'utiliser crête comme une caractéristique pour la représentation d'objet. Cette proposition n'est pas une exclusion des caractéristiques classiques telles que le contour, la région, le squelette. Nous reconnaissons que l'intégration des caractéristiques (contour, crête, pic, point d'intérêt) peuvent produire des améliorations pour la reconnaissance, mais comment peut on les combiner reste une question très difficile. Cette thèse ne s'intéresse qu'aux crêtes pour la représentation d'objets.

La figure 1.9 montre les crêtes détectées de l'image du vélo à l'échelle $\sigma = 4\sqrt{2}$. En comparaison avec les caractéristiques telles que la région, le contour ou les points d'intérêt, les crêtes représentent le plus significativement les structures du vélo. Les structures allongées sont représentées par les crêtes longues tandis que les structures plutôt courtes sont représentées par une crête courte qui peut être considérée comme un pic.

centrées dans la forme d'origine.

[8] Sa définition est présentée dans le chapitre 3. Nous allons y voir pourquoi les crêtes sont appelées les lignes d'intérêt naturelles.

1.2 Reconnaissance d'objets

La reconnaissance d'objets se décompose en deux phases :

– Une phase de modélisation et d'apprentissage permet d'associer un modèle à chaque classe d'objet. Cette phase construit une base de modèles qui est l'unique représentation des objets pour la deuxième phase.

– La phase de reconnaissance consiste à mettre en correspondance une image avec un élément de la base de modèles. Cet appariement mesure le degré d'appartenance de l'image au modèle.

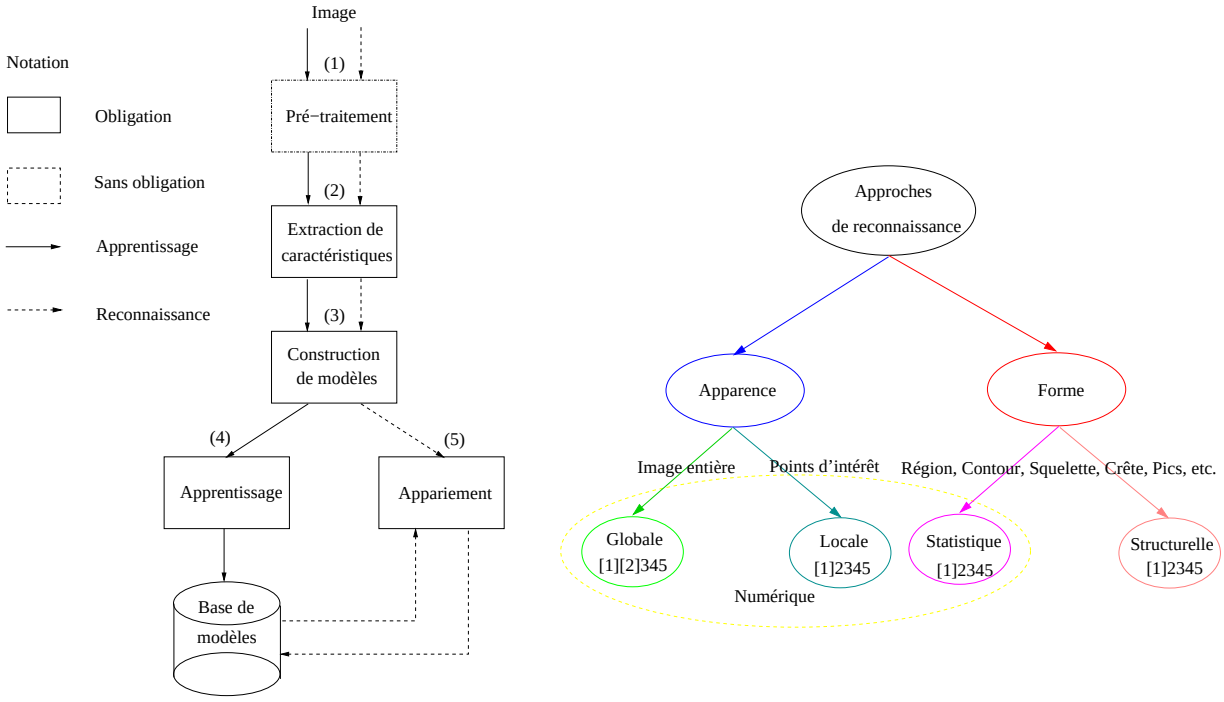


FIG. 1.10 – (a) Composantes principales d'un système de reconnaissance d'objets. (b) Classification de méthodes de reconnaissance d'objets.

La figure 1.10a montre les étapes d'un processus d'apprentissage et de reconnaissance. Deux étapes (1) et (2) ne sont pas obligatoires. Par exemple, les approches globales basées sur l'apparence ne nécessitent pas l'extraction de caractéristiques.

1.2.1 Evolution des méthodes de représentation

La construction des modèles d'objet est au coeur d'un système de reconnaissance. Les recherches sur la modélisation d'objets peuvent être classifiées en deux catégories. La première catégorie contient des approches de représentation basées sur la forme [Cro81, MA88, TC04, BL93, PSZ99, SKK01, SSDZ98, RM93, KD05]. La deuxième catégorie contient des approches basées sur l'apparence [SK87, Swa91, MN95, PPP98, SC96a, SC96b, SC00, dV99, HdVC00, HCCdV00, Hal01, SM97]. Ces deux

types d'approche ont été combinés pour construire des modèles plus robustes [OFPA04a, MHSTS05, VJ01, OFPA04b, TMF04].

Il est intéressant de noter que depuis 30 ans, les méthodes de reconnaissance d'objets ont évolué en amenant le modèle près de l'image. Plus concrètement, au début les auteurs construisent la représentation basée sur la forme géométrique, une propriété intrinsèque de l'objet. Peu à peu, les auteurs modélisent l'objet en le rendant moins dépendant de sa géométrie. Actuellement, la plupart des approches se basent sur l'apparence de l'objet. Ces approches sont indépendantes de la géométrie de l'objet. Quelle est l'origine de cette évolution ?

Dans des années 70, les chercheurs ont représenté un objet par ses composantes volumiques, par exemple les cylindres généralisés [Bro83] ou plus tard les superquadriques [Pen86] et les géons [Bie85]. Un grand défi de ces systèmes est le fossé entre les caractéristiques qui devraient être détectées fiablement et la nature abstraite des composantes.

Les années 80 ont fait émerger une approche de modélisation des objets qui cherche à capturer exactement la forme géométrique de l'objet. Ces modèles sont souvent des modèles CAO [Low85, HU90]. Cette fois, le fossé est réduit en amenant le modèle proche de l'objet. Pourtant, il demande de connaître à priori la géométrie de l'objet. De plus, comme la présence de texture affecte sérieusement la complexité de tels systèmes, ils ont du sélectionner les objets simples tels que les objets manufacturés. En réalité, ces systèmes n'ont pas été capables de reconnaître des objets complexes.

Pour réduire le fossé entre "observables" et "représentables", plusieurs travaux des années 90 ont concentré à porter l'image près de modèle en éliminant des surfaces de représentation, contrôlant la condition de lumière et réduisant l'encombrement de fond. Pour cela, l'image est représentée par un ensemble de mesures numériques calculées sur l'image entière ou quelques points d'intérêt dans l'image. Ces mesures sont facilement adaptées aux paramétrages.

Les premières approches de reconnaissance d'objets basées sur l'apparence ont beaucoup de difficultés face aux problèmes tels que l'occultation, la rotation, la translation, le changement d'échelle, l'encombrement du fond. Les approches actuelles ont essayé de réduire ces limitations en calculant les mesures à l'échelle appropriée et les normaliser par l'énergie moyenne, etc. De cette manière, ces approches ont obtenu des résultats remarquables de reconnaissance de différents types d'objets. Elles attirent de plus en plus la popularité de la communauté de vision par ordinateur.

Malgré l'excellence des méthodes basées sur l'apparence, il est important de noter qu'en amenant le modèle d'objet près de l'image, les méthodes s'éloignent de l'objectif de la reconnaissance générique où on a besoin d'une caractérisation abstraite de la forme. Ainsi, l'hypothèse que ces méthodes peuvent résoudre le problème de reconnaissance prototypique apparaît incertaine.

En fait, toutes les méthodes de reconnaissance basée sur l'apparence sont fondées sur une hypothèse qu'il y a une correspondance entre une caractéristique (une ligne de contour, une région entourée d'un point d'intérêt, etc) avec une caractéristique dans le modèle. Dans le cas de reconnaissance générique, cette supposition n'est pas toujours vraie parce qu'il se peut qu'une saillance dans une image n'implique pas une saillance dans le modèle. La reconnaissance générique d'objet devrait générer les abstractions qui peuvent ne pas explicitement apparaître dans l'image mais qui capturent la saillance de la catégorie.

Dans les sections ci-dessous, nous résumons quelques méthodes de représentation. Nous les divisons en deux catégories : la première catégorie contient des méthodes basées sur l'apparence et la deuxième catégorie contient des méthodes basées sur la forme. Chaque catégorie peut être divisée encore selon la technique de représentation (numérique ou structurelle). L'exposé des méthodes suit la classification illustrée dans la figure 1.10b.

1.2.2 Représentation d'objets basée sur l'apparence

L'utilisation de modèles de type CAO souffre d'un problème fondamental : on ne sait pas s'il existe de bons algorithmes pour trouver dans une image les caractéristiques constituant le modèle ; par exemple, il s'est avéré impossible de trouver correctement des arêtes de contours, des faces planes ou cylindriques. La vision par apparence[9] prend une approche résolument opposée : on construit les modèles à partir d'images, quitte à ne pas avoir d'interprétation géométrique ou intuitive des caractéristiques utilisées. C'est donc une approche éminemment pragmatique.

La reconnaissance d'objets par apparence repose essentiellement sur des techniques numériques. Chaque composante de l'objet est décrit par des descripteurs. L'objet est alors un nuage de points dans l'espace des descripteurs, espace qui est souvent représenté par un histogramme ou une Gausienne multi-dimensionnelle. Dans cette représentation, aucune information structurelle de l'objet n'est explicite.

Les approches de modélisation d'objet basées sur l'apparence diffèrent selon le nombre de vecteurs de descripteurs utilisés pour décrire une image. Les approches globales utilisent l'image entière pour l'encoder en un seul vecteur de descripteur tandis que les approches locales calculent plusieurs vecteurs de descripteurs sur les imagettes dans l'image.

Approches globales

Les approches globales considèrent l'ensemble de tous les pixels dans l'image. Aucune hypothèse n'est faite sur la position et la présence des éléments dans l'image. Ce type d'approche traite toutes les images de manière équivalente et ne dépend pas de descripteurs spécifiques. Quelques approches utilisent l'image originale comme modèle, quelques autres se basent sur les techniques statistiques telles que l'histogramme de couleur ou de texture.

L'avantage commun des approches globales est la simplicité. Pourtant, en utilisant l'image entière pour la modélisation, elles ne permettent pas une variation de l'apparence des objets dans une catégorie y compris l'erreur de normalisation, l'occultation partielle et le changement de luminosité.

1. *Description de l'image par la technique ACP*

   La façon la plus simple de reconnaître un objet dans l'image est de réaliser un appariement direct d'images. Cet appariement utilise une mesure de corrélation entre images pour mettre en correspondance une image avec un modèle. Différentes techniques de corrélation peuvent être utilisées suivant des conditions expérimentales (voir [MC95]).

   L'appariement d'images est approprié pour comparer deux images ou déterminer si un objet dans l'image se trouve également dans une autre. Néanmoins, cette approche ne peut être appliquée pour de nombreux objets sous multiples points de vue parce qu'elle demande une énorme quantité de mémoire pour stocker des modèles et un temps de recherche considérable. Pour réduire la mémoire de stockage et le temps de recherche, quelques auteurs ont proposé de convertir l'espace d'images en un sous-espace des images propres de faible dimensionnalité.

   La technique statistique de l'ACP (Analyse en Composantes Principales) est une technique permettant de déterminer un sous-espace optimal pour la représentation d'une quantité importante de données. Cette technique a été utilisée en vision par ordinateur pour reconnaître des visages [SK87],

[9] appearance based vision

reconnaître des objets rigides [PPP98], compresser des images vidéos [CCBS97, CdVC99] ou estimer la position d'un robot mobile [Pou98, WSV99].

Pour la reconnaissance d'objets, dans [MN95], Murase et Nayar ont proposé de construire une base de descripteurs en se basant sur les vecteurs propres correspondant aux plus grandes valeurs propres de la matrice de covariance d'une matrice définie par la concaténation des images d'apprentissage. La dimension de ce sous-espace réduit significativement par rapport à celle de toutes les images. Après la construction de l'espace des images propres, les images d'apprentissage y sont projetées et les vecteurs de mesures seront stockés comme modèles.

La reconnaissance d'un nouvel objet consiste à segmenter l'objet de l'image, normaliser la taille et la luminosité, projeter dans l'espace des images propres apprises, puis comparer des vecteurs de mesures de faible dimensionnalité. Cette approche obtient un taux de reconnaissance de l'objet 3D très élevé en supposant que les objets ne sont pas occultés et la segmentation de l'objet est faisable (100% si le nombre de dimensions de sous-espace est suffisamment grand ≥ 6). Dans un cas général, cette technique est sensible au bruit, au déplacement de l'objet, au changement d'échelle et de luminosité. Elle est plus appropriée à la reconnaissance de la pose et l'orientation de l'objet que la discriminance des classes d'objets.

2. *Approches basées sur la couleur*

   La couleur est une caractéristique distinctive des objets. Dans la nature, il y a plusieurs objets qui sont très similaires en forme et ne sont distingués que par la couleur comme par exemple le cas d'une pomme et une tomate. Les premiers systèmes d'indexation d'images ont modélisé chaque image par un histogramme de couleur dans l'espace RGB [Swa91]. Cet espace de couleur est bien adapté à l'affichage sur l'écran mais loin de la perception humaine. Dans l'optique de modéliser l'objet par la couleur perceptuelle, d'autres espaces de couleurs ont été expérimentés comme par exemple, L*u*v, L*a*b.

   La reconnaissance d'objets s'effectue en comparant l'histogramme de couleur de nouvel objet avec les histogrammes des objets appris. La technique de modélisation d'objet par l'histogramme de couleur, comme montrée par Swain et Ballard dans [Swa91], est remarquablement robuste au changement de l'orientation, changement de l'échelle de l'objet. Pourtant, une faiblesse de cette approche est la sensibilité au changement d'éclairage, la pose et le point de vue. Par exemple, pour les images d'extérieur, la couleur d'éclairage varie selon le temps et les conditions atmosphériques. La couleur d'un objet dans une scène peut également varier en fonction des ombres, de l'interéflexion avec d'autres objets. Ainsi, la couleur aide à reconnaître des objets dans un cas où la forme n'est plus discriminante (ie. le cas de tomate et pomme). Mais seule, elle ne garantit pas une reconnaissance fiable.

   Dans [LS03], Leibe et Schiele ont montré des résultats de reconnaissance de classes d'objets par la couleur. Les objets tels que la tomate, la pomme, le poivre, le buffle, le chien, le cheval, le verre, la voiture "jouet"[10] sont acquis dans des conditions idéales et tous sont segmentés manuellement pour que le fond n'intervienne pas dans la reconnaissance. L'histogramme de couleur donne un taux de reconnaissance le plus bas par rapport aux méthodes basées sur le texture ou la forme (64.85% en moyen). Dans un environement réel où les objets sont plus complexes et la segmenta-

[10] Les images se trouvent sur le site http ://www.vision.ethz.ch/pccv/

tion n'est pas parfaite, l'approche basée sur la couleur ne dépasse pas ce taux de reconnaissance.

Pour améliorer la performance de la méthode basée sur la couleur, quelques auteurs ont introduit les mesures moins sensibles au changement de la lumière. Par exemple, Funt et Finlayson ont proposé d'utiliser les dérivées de logarithmes de couleur pour fournir une invariance à la couleur [FF95]. Healey et Slater utilisent les moments de l'histogramme de couleur pour augmenter la robustesse au changement de l'intensité de lumière [HS94]. Hunke [Hun94] et Freeman [FA91] ont montré que la normalisation du vecteur de couleur par luminance fournit un moyen fiable pour détecter la couleur de peau pour le suivi de visage.

3. *Approches basées sur la texture*

   Les approches basées sur la texture sont en fait une extension des approches basées sur la couleur. Au lieu d'utiliser l'histogramme de couleur, les auteurs utilisent l'histogramme de dérivées du signal d'image. Ainsi, la reconnaissance consiste à comparer de deux histogrammes de dimensionnalité multiple.

   Dans [SC96a, SC96b, SC00], Schiele et al. ont proposé d'utiliser l'histogramme de champs réceptifs. Cette représentation comme toutes les méthodes basées sur la technique de l'histogramme n'a pas besoin de segmentation ni de modélisation géométrique de l'objet. Les champs réceptifs sont en fait la première dérivée de Gaussienne, la magnitude et direction de la première dérivée et le Laplacien. L'effet de variation de l'intensité du signal est éliminé par la normalisation des réponses des filtres de champs réceptifs par l'énergie du signal. Différentes combinaisons de filtres ont été expérimentées sur la base Colombia. Avec le test χ^2, le taux de la reconnaissance est de l'ordre 100% qui montre une excellente performance de la méthode, malgré un changement de point de vue et un léger changement de l'échelle.

   Dans un récent article [LS03], Leibe et Schiele ont ré-expérimenté l'approche de représentation basée sur l'histogramme de dérivées de Gaussiennes D_x et D_y sur une nouvelle base d'images crée par eux-même pour une comparaison des méthodes de reconnaissance. Cette technique a obtenu un résultat parfait de reconnaissance des objets dans la base Colombia. Dans le nouveau test, seulement 80% des objets ont été reconnus correctement. La raison est que la plupart des objets dans la base Colombia sont texturés tandis que la majorité des objets dans la nouvelle base n'ont pas de cette caractéristique.

Approches locales

Les techniques de représentation basées sur l'apparence utilisent l'information statistique sur l'image entière, donc intrinsèquement globale. Comme tous les pixels dans l'image sont considérés, il se peut que l'image soit caractérisée plus par les pixels du fond que par les pixels de l'objet d'intérêt. Pour permettre une modélisation plus correcte de l'objet, ce dernier doit être nettement segmenté et normalisé. Or la segmentation correcte est coûteuse et très difficile voir impossible à réaliser. Dans un environnement complexe, un objet peut apparaître partiellement occulté, la technique globale a alors beaucoup de difficultés à le reconnaître.

Pour remédier au problème de l'occultation, les approches locales basées sur l'apparence sont proposées. Le principe des méthodes locales est de décomposer l'image originale en plusieurs imagettes de petite taille. La mise en correspondance ne sera pas effectuée entre les images entières mais entre les

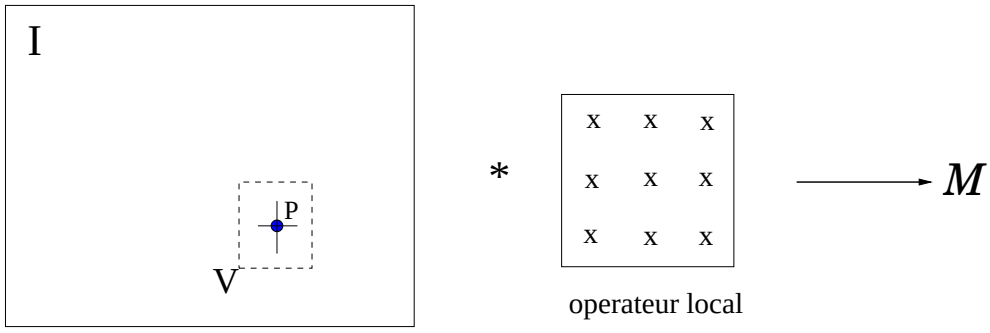


FIG. 1.11 – L'opérateur local appliqué sur un voisinage d'un point dans l'image fournit une mesure du voisinage.

imagettes. De cette manière, elle assure que dans toute nouvelle image, une partie de l'imagette sera présente donc reconnaissable.

Le problème de reconnaissance basée sur l'apparence locale peut être divisé en 3 sous-problèmes. Le premier consiste à définir les régions auxquelles les descripteurs seront appliqués. Le deuxième est de choisir l'ensemble de descripteurs à appliquer. Le troisième est de définir une structure de représentation optimale pour le stockage et efficace pour la recherche. La reconnaissance s'effectue soit par l'appariement des caractéristiques locales [OI96, OI97, Sch92] (dans ce cas, une technique de vote ou de transformée de Hough peut être utilisée directement pour sélectionner l'objet le plus vraisemblable), soit par l'appariement des graphes de caractéristiques locales [dV99] ou la statistique sur les caractéristiques locales [Swa91, SC96c].

Les différences entre les approches existantes viennent de la sélection de base de descripteurs ou de positions dans l'image pour appliquer les descripteurs choisis. Concernant la sélection des positions pour appliquer les descripteurs, il y deux tendances : l'une applique l'opérateur de descripteur sur tous les points, l'autre applique seulement sur quelques points spéciaux comme les corners, les jonctions, les pics, etc. Les descripteurs caractérisant la structure locale en un point sont les descripteurs obtenus par la technique PCA [dV99] ou les filtres de Gaussiennes [dV99, HdVC00, HCCdV00, Hal01, SM97].

Un descripteur local est un opérateur dont le support spatial est faible par rapport à la taille de l'objet (moins de 5% de la taille de l'image). L'intérêt de ce caractère local consiste à n'utiliser qu'une faible partie de l'image pour évaluer les vecteurs de mesures. Ainsi, il permet à la reconnaissance de tolérer l'occultation partielle. La figure 1.11 illustre l'application d'un opérateur local sur un voisinage V correspondant au point P(x,y) dans l'image. Le résultat de cette application est un scalaire M, appelé la mesure de l'imagette.

1. *Descripteurs locaux basés sur la technique ACP*

   La première base de descripteurs locaux est construite à partir d'une extension directe de la technique ACP proposée par Ohba et IKeuchi [OI96, OI97]. Au lieu de calculer un espace propre des images, ils calculent un espace propre des imagettes de celles-ci. Cette technique s'appelle *technique de fenêtres propres*. La taille des imagettes utilisées dans leur expérimentation est 15x15, ce qui forme une sous-espace de 255 dimensions. Cette réduction de dimension permet un calcul plus rapide.

Une contribution importante dans l'approche de Ohba et Ikeuchi est qu'ils ont proposé des critères pour sélectionner les imagettes optimales. Il s'agit des imagettes qui satisfont des critères de *détectabilité, unicité et fiabilité* dans l'espace propre. La sélection de imagettes basée sur ces critères permet de réduire l'espace de mémoire utilisée pour stocker les modèles. En outre, les modèles sont plus discriminants.

L'approche de Ohba et Ikeuchi profite de la réduction de la dimension de l'espace de descripteurs par la technique ACP. Elle est invariante à la position. De plus, grâce à la description de caractéristiques locales, l'approche est moins sensible à l'occultation.

L'espace propre que construisent Ohba et Ikeuchi est un espace de niveau de gris. Pour augmenter la discriminance des données, qui apparaissent parfois trop importantes, Colin de Verdières propose d'utiliser l'information de couleur [dV99]. Plus spécifiquement, il concate 3 composantes de couleurs R, G, B dans la construction de la matrice des images. Ajouter l'information de couleur augmente la dimension de l'espace propre et donc rend plus discriminante la reconnaissance. Leurs expérimentations sur une base de 1000 images montrent que la chrominance permet de réduire considérablement les cas d'échec (25% en cas de luminance et seulement 10% en cas de chrominance).

Dans [dV99], Colin de Verdière a montré que la reconnaissance d'objets par les descripteurs construits par la technique ACP est satisfaisante. Pourtant, cette base de descripteur est restreinte à la reconnaissance d'objets vu sous un point de vue similaire à l'un des points de vue d'apprentissage. Sans apprentissage à orientations et échelles variables, la reconnaissance ne peut être robuste à ces paramètres que de façon très limitée. Pour résoudre cette limitation, une base de descripteurs formés à partir des dérivées de Gaussiennes est proposée.

2. *Descripteurs locaux basés sur les filtres de Gaussiennes*

   Pour avoir une représentation invariante à l'échelle, il faut choisir une base de descripteurs qui approche aussi précisément que possible les imagettes observées par un nombre de descripteurs aussi faible que possible [dV99] en permettant de paramétrer l'échelle facilement. Une approximation classique de Taylor permet une représentation du signal en un point par l'ensemble des vecteurs de dérivées appelé "Jet Local" par Koenderink [KvD87].

   Rao [RB95] propose de représenter un objet par un ensemble de vecteurs de mesures locales fondées sur des descripteurs dérivées de Gaussiennes. Le vecteur est constitué de 45 dimensions à 5 échelles et 9 dérivées reparties sur les ordres 1, 2, 3. La modélisation est réalisée par l'enregistrement des vecteurs dans une base puis la reconnaissance s'effectue par la recherche du vecteur similaire. Cette approche basée sur les dérivées normalisées en orientation est ainsi invariante à l'orientation 2D. De plus, le nombre de dimensions élevé permet de bien discriminer des vecteurs, très peu de faux appariements ont été rencontrés.

   Le fait de calculer des mesures sur les imagettes déterminées en tous les pixels de l'image est coûteux et parfois inutile car deux imagettes similaires donnent une mesure similaire. Ainsi, dans [Sch92], Schmid a proposé de calculer les mesures seulement sur les points renvoyés par le détecteur de Harris modifié. Chaque imagette est caractérisée par un vecteur de combinaisons de dérivées de Gaussiennes. Ce vecteur est invariant à la rotation rigide de l'image. La reconnaissance s'effectue par un algorithme de vote suivant la mise en correspondance des multiples vecteurs de

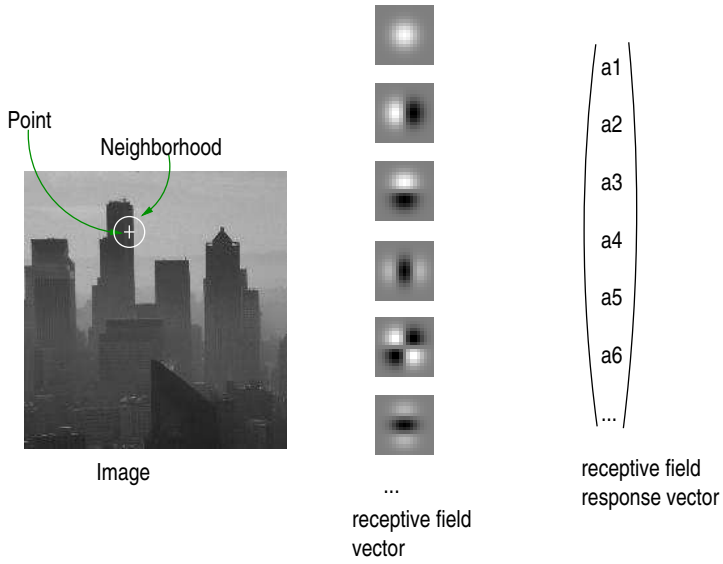


FIG. 1.12 – Un exemple de représentation d'une imagette par un vecteur des champs réceptifs.

descripteurs.

Colin de Vedière [dV99] propose une combinaison de 9 composantes pour mesurer une imagette : $M(x,y) = [L_0^1, L_{\frac{\pi}{2}}^1, L_0^2, L_{\frac{\pi}{3}}^2, L_2^2, L_0^{\frac{2\pi}{2}}, L_{\frac{\pi}{4}}^3, L_{\frac{\pi}{2}}^3, L_{\frac{3\pi}{4}}^3]^T$ ou L_θ^i est la dérivée d'ordre i selon la direction θ. Ces dérivées sont calculées aux positions de pics détectés selon Lindeberg [Lin94]. Pour rendre la méthode invariante à l'échelle, l'orientation et la luminance, il sélectionne l'échelle optimale de pic, projette les composantes à cette échelle, puis calcule l'orientation du vecteur Gradient à cette échelle et projette les composantes à cette orientation.

La description de l'image selon Colin de Verdière, est une surface 2D (une grille 2D en fait) construite par la projection des imagettes sur l'espace de descripteur de 8 dimensions (la deuxième composante projetée à l'orientation du Gradient s'annule). La reconnaissance s'effectue en mettant en correspondance une nouvelle surface avec une surface apprise par un algorithme de vote. Pour restreindre la mise en correspondance, Colin de Verdière a ajouté la contrainte de la position relative des imagettes. L'algorithme permet, simultanément, l'identification de l'objet et la détermination de sa pose dans l'image.

Hall [Hal01] a observé que l'ajout de l'information de couleur dans la représentation proposée par Colin de Verdière permet de rendre la représentation plus discriminante et ainsi que la reconnaissance. Ainsi, dans sa thèse, elle a représenté une imagette par un vecteur de descripteurs de champs réceptifs des couleurs : $M(x,y) = [L_0^{1Y}, L_1^C, L_2^C, L_0^{1C_1}, L_0^{1C_2}, L_0^{2Y}, L_{\frac{\pi}{4}}^{2Y}, L_{\frac{\pi}{2}}^{2Y}]^T$ où Y, C_1, C_2 sont les canaux de couleur dans l'espace YC_1C_2. Cet espace de couleur a été choisi de façon appropriée pour séparer la luminance et la chrominance. Ainsi, la description capture à la fois l'information de luminance et de texture.

Le résultat de reconnaissance par la méthode proposée par Hall est expérimentalement meilleur

que celui des deux méthodes originales basées seulement sur la couleur [Swa91] ou les dérivées de l'achromiance [dV99]. Une amélioration considérable du taux de reconnaissance d'un coefficient de 0.70 (par rapport 0.54 pour la méthode basée sur la couleur et 0.16 pour la méthode basée sur les champs réceptifs) dans le cas d'objet sur un fond complexe montre une amélioration de la performance de la méthode.

On peut observer que l'utilisation d'un seul type de point d'intérêt et le calcul des descripteurs seulement sur les imagettes correspondant aux points d'intérêt ne permettent pas de capturer tous les aspects de l'objet. Dans [VJ01, AR02, OFPA04a, OFPA04b], les auteurs ont proposé de combiner des vecteurs de descripteurs calculés dans les imagettes correspondant aux plusieurs types de points d'intérêt : "Harris-Laplacian", "SIFT"[11], "Salient Points", etc. Dans [OFPA04a], Opelt *et al.* ont calculé les descripteurs non seulement sur les zones avec un changement fort du signal, mais aussi sur les régions uniformes. Le modèle de l'objet est une probabilité des combinaison des mesures calculées à partir de ces imagettes. Avec la combinaison de différentes mesures sur différents types d'imagettes, le résultat de la reconnaissance a été amélioré de façon significative.

Description détaillée de quelques méthodes locales de *reconnaissance de classes d'objets* basées sur l'apparence

Les approches globales ou locales de reconnaissance basées sur l'apparence présentées ci-dessus ne s'inscrivent pas dans le contexte de *la reconnaissance de classes d'objets*. Elles sont plutôt pour la détermination de la pose et de l'orientation de l'objet dont le modèle est connu. Cette sous-section décrit en détail quelques approches locales proposées pour reconnaître des classes d'objets basées sur l'apparence. Ces approches considèrent que l'objet se constitue d'un certain nombre de composantes significatives.

Agarwal et Roth ont proposé une approche simple de détection automatique des objet dans une image en apprenant les composantes significatives de ceux-ci [AR02]. L'approche consiste en 4 étapes :

1. *Construction de vocabulaire* : Pour obtenir une représentation de classes d'objets d'intérêt, les auteurs utilisent l'opérateur de point d'intérêt de Förstner pour sélectionner les composantes distinctives de la classe d'objet. Il s'agit des régions d'intersection de deux lignes ou des régions circulaires. Des régions de taille fixe (ie. 13x13) autour des points d'intérêt sont obtenues (figure 1.13). Le vocabulaire est un ensemble de régions indexées par des indices. Les composantes similaires sont assignées par un même indice.

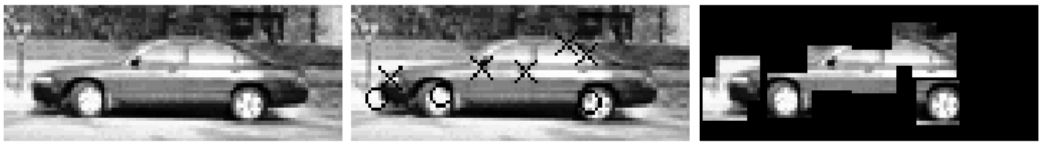


FIG. 1.13 – De gauche à droite : Image originale, Les points fournis par le détecteur Förstner. Les composantes correspondant aux points d'intérêt sont coupées de l'image originale.

[11] Scale Invariant Feature Transform

2. *Représentation de l'image* : Ayant construit le vocabulaire, la représentation d'images est effectuée en déterminant quelles composantes sont présentes dans l'image. Ceci est réalisé par la détection des points d'intérêt pour sélectionner des composantes puis comparer les composantes avec celles du vocabulaire construit. Si une similarité est obtenue, la composante considérée est assignée au même indice que celui de la composante similaire dans le vocabulaire. L'image est donc représentée simplement par un vecteur binaire où la valeur 1 à la position i indique la présence de la composante i dans l'image et 0 son absence.

   Pour augmenter la discriminance entre de différents objets, les relations spatiales entre les composantes sont prises en compte. Chaque paire de composantes est caractérisée par la distance et la direction entre les composantes. Dans leur expérimentation, l'espace de distance est discrétisé en 5 et l'espace de direction est discrétisé en 4. Ceci donne 20 combinaisons possibles de (distance, orientation). Chaque relation est numéroté par un nombre de 1 jusqu'à 20. L'image est alors représentée par un vecteur de 270 éléments binaires qui sont les indices des composantes et celles de relations. Parmi 270 éléments, il y a très peu d'éléments actifs. Cette représentation est ainsi appelée représentation éparse.

3. *Apprentissage de classifieurs* : Etant donnés des exemples positifs (objet) ou négatifs (non-objet), chaque image d'exemple est représentée par un vecteur binaire comme décrit ci-dessus. Ces vecteurs sont ensuite donnés à l'entrée d'un algorithme d'apprentissage supervisé SNoW (Sparse Network of Winnows [KV99]) qui apprend à classifier une image comme membre ou non-membre d'une classe d'objet.

4. *Détection de l'objet dans une nouvelle image* : Etant donnée une nouvelle image, en déplaçant une fenêtre de taille 100x40, un vecteur binaire est calculé et classifié comme étant positif ou négatif. Il est possible que plusieurs fenêtres représentent le même objet. Pour éliminer le risque de surdétection, les auteurs ont proposées une carte d'activation qui permet de marquer des activations et de choisir la fenêtre ayant la plus grande valeur d'activation comme étant la meilleure localisation de l'objet détecté.

L'approche de Agarwal et Roth a été expérimentée pour la détection de voitures dans une image et a obtenu un résultat de détection satisfaisant : les meilleurs rappel et précision sont respectivement de 81% et 77%. Un désavantage de l'approche est la limitation au changement de la taille de l'objet dans l'image. Les raisons de ceci viennent de la détection des points d'intérêt sans adaptation à l'échelle, puis de l'utilisation de la taille et l'orientation fixes pour caractériser les composantes. En outre, dans cette approche, la vérification est coûteuse parce que toutes les relations entre deux composantes doivent être considérées.

Pour remédier au problème de changement d'échelle, **Hall** a proposé une approche similaire de détection de classes d'objet dans une image [Hal04]. Pourtant, dans son approche, les détecteurs sont choisis indépendamment les uns des autres et l'organisation spatiale est représentée par un graphe élastique qui permet de localiser la correspondance des composantes de façon plus précise. Cette approche de reconnaissance se compose de 4 étapes :

1. *Sélection des détecteurs* : Chaque point de l'image d'apprentissage est représenté par un vecteur de champs réceptifs d'ordre 1 et 2 (donc un vecteur de 5 éléments). Les champs réceptifs de luminance sont calculés à l'échelle intrinsèque du point considéré qui est l'échelle à laquelle la valeur de Laplacien est extremale. Tous les points dans toutes les images d'apprentissage fournissent

un ensemble de vecteurs de champs réceptifs. Ces vecteurs sont les points dans un espace de 5 dimensions. Pour associer les points proches dans l'espace de caractéristiques, un algorithme de K-Means est appliqué et le nuage de points l'espace est un mélange des Gaussiennes. Alors, les détecteurs sont les clusters dans l'espace de caractéristiques. Le nombre de clusters est le nombre de détecteurs.

FIG. 1.14 – Gauche : Image originale. Droite : Carte d'indices des clusters en chaque point dans l'image

2. *Représentation de l'image* : Etant donnée une image, pour chaque point de l'image, son échelle caractéristique est déterminée et le vecteur de champs réceptifs est calculé. Ce vecteur est projeté dans l'espace de caractéristiques et le point est assigné à un indice qui est l'indice de cluster duquel le vecteur est le plus proche. Ce processus est répété pour tous les points dans l'image et le résultat obtenu est une carte des indices.
3. *Appariement élastique des graphes étiquetés* : Les objets modèles sont représentés par des graphes dont les noeuds sont assignés par une description de l'apparence locale (le vecteur de champs réceptifs) et les arcs sont étiquetés par la distance entre deux noeuds. Le problème de la détection d'un nouvel objet dans une image, étant donné un objet modèle, revient au problème de mise en correspondance de deux graphes étiquetés. Ceci est un problème d'optimisation d'une fonction de coût qui combine la similarité entre les noeuds et la similarité géométrique du graphe. Pour cela, un graphe rigide est trouvé à la meilleure position en balayant le graphe avec un pas suffisamment grand. Ensuite, les noeuds dans le graphe sont étirés pour obtenir la meilleure correspondance entre des noeuds.

L'approche proposée par Hall a plusieurs avantages. D'abord, les détecteurs sont choisis de façon appropriée à la classification. Elle ne dépend pas de détecteurs comme quelques approches proposées par Agarwal et Roth, Fergus et al. Ensuite, les descriptions locales sont invariantes à l'échelle. Ainsi, cette approche est robuste au changement de l'échelle. Enfin, l'algorithme d'appariement des graphes élastiques est efficace en terme de temps de calcul. La flexibilité des noeuds dans un graphe permet une distorsion locale des composantes de l'objet dans l'image. L'approche est expérimentée pour 4 catégories d'objets : visage, moto, avion, voiture vue de derrière. Un bon résultat a été obtenu.

Malgré les avantages, on trouve quelques inconvénients de l'approche. D'abord, les descripteurs ne sont pas invariants à l'orientation. Ensuite, la taille ainsi que le nombre de noeuds dans le graphe sont des paramètres choisis manuellement. La position du graphe modèle dans l'image modèle doit être

déterminée à la main ce qui rend l'approche appropriée à la recherche d'images (comme par exemple Google Image) plutôt qu'à la reconnaissance générique d'objets.

Récemment, **Fergus et al.** ont proposé une approche de représentation des classes d'objets, qui a fait un grand éclat dans le domaine de la reconnaissance. Cette approche combine l'information de la forme et de l'apparence de l'objet dans un modèle probabiliste, comme développé par Burl, Weber et al. [BWP98, WWP00b, WWP00a]. Nous expliquons le principe de la méthode ci-dessous :

1. *Détection de caractéristiques* : Les caractéristiques sont détectées par le détecteur de points saillants proposés par Kadir et Brady [kB01]. Les points saillants sont des points qui maximise l'entropie locale en ce point (figure 1.15). Ils sont détectés de façon invariante à l'échelle.

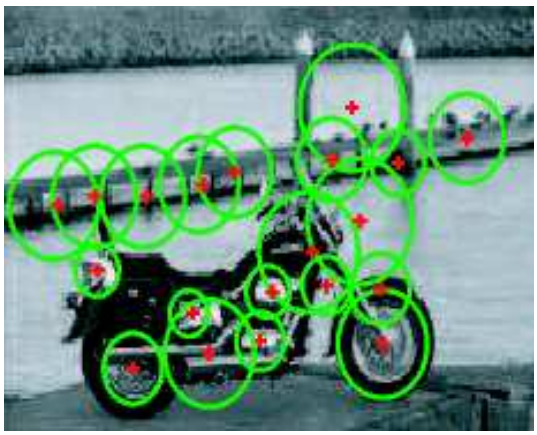


FIG. 1.15 – Quelques points les plus significatifs détectés par le détecteur de Kadir et Brady.

2. *Représentation de caractéristiques :* Chaque point saillant détermine une région d'intérêt. La région est décrite par la position du centre X et l'échelle du point S considéré. Chaque région est coupée de l'image, normalisée à la taille 11x11 et représenté par un vecteur de 121 composantes qui sont des informations monochromes. Les images d'apprentissage fournissent l'ensemble de tels vecteurs, qui sont ensuite entrés dans un algorithme d'analyse en composantes principales. Chaque vecteur de 121 composantes est transformé en un point A dans l'espace à 15 dimensions qui correspondent à 15 valeurs propres les plus grandes. Finalement, chaque caractéristique est représentée par un triplet (X, S, A).
3. *Apprentissage :* L'apparence d'une caractéristique est un vecteur de 15 dimensions. Chaque caractéristique a une densité Gaussienne, dont la moyenne et la covariance sont ${\theta_p}^{app} = \{c_p, V_p\}$. L'apparence de fond a une densité Gaussienne ${\theta_{bg}}^{app} = \{c_{bg}, V_{bg}\}$.

   La forme est représentée par une densité Gaussienne jointe des locations des caractéristiques. Elle a les paramètres $\theta^{shape} = \{\mu, \Sigma\}$.

   L'échelle est aussi modélisée par une densité gaussienne $\theta^{scale} = \{t_p, U_p\}$.

   La tache d'apprentissage consiste à estimer les paramètres $\theta = \{\mu, \Sigma, c, V, M, p(d|\theta), t, U\}$. Le but est de trouver les paramètres qui maximisent la vraisemblance : $\theta_{ML} = argmax\ p(X, S, A|\theta)$. Ceci s'effectue par une technique EM.

4. *Reconnaissance :* La reconnaissance consiste à détecter des caractéristiques dans la nouvelle image, et les représenter par un triplet (X, S, A). En calculant le ratio de vraisemblance R et en seuillant cette valeur, la présence ou l'absence d'un objet dans une image sera déterminée.

L'approche basée sur le modèle probabiliste a obtenu un excellent résultat de catégorisation d'objets : le taux d'erreur est de 10%. Cette approche a été utilisée comme moteur de recherche des images "Google Images" [FPZ04].

Un défaut de l'approche est qu'elle dépend trop du résultat de détection de caractéristiques. Si le détecteur de point saillant de Kadir et Brady ou n'importe quel détecteur utilisé ne donne pas de bonnes localisations des composantes souhaités, l'approche échoue. En outre, la complexité de l'approche augmente exponentiellement avec le nombre de caractéristiques significatives utilisées pour modéliser l'objet. Ainsi, le nombre de caractéristiques est limité (≈ 30). A cause de cette limitation, quelques composantes intéressantes de l'objet peuvent être ignorées.

Toutes les approches présentées ci-dessus ne sont pas invariantes à la transformation affine. En fait, les détecteurs utilisés pour la détection des points d'intérêt ne le sont pas. Dans [MS04b, MS04a, MS05], Mikola et Schmid ont adapté le détecteur Harris-Laplacien pour qu'il soit invariant à la transformation affine. Ce détecteur est utilisée dans le travail très récent de **Dorko et Schmid** pour détecter des composantes invariantes à l'échelle et à la transformation affine [DS05]. Ces composantes sont apprises par un mélange de Gaussiennes et utilisées pour la classification d'objets.

1. *Sélection de caractéristiques* : Trois détecteurs de points significatifs sont utilisés pour détecter des caractéristiques : Détecteur de points saillants de Kadir et Brady [kB01], Détecteur de points d'intérêt adapté à l'échelle Harris-Laplacien de Mikola et Schmid [MS01], Détecteur de points invariants à la transformation affine Harris-Laplacien-Affine [MS04b, MS04a, MS05].

   Chaque point détecté détermine une région circulaire ou elliptique (figure 1.16). Toutes les régions correspondant à des points obtenus sont normalisées en régions circulaires. L'invariance par rapport à la rotation est obtenue en tournant la région circulaire en direction du vecteur Gradient moyen calculé sur une petite région au tour du point considéré. Chaque région circulaire est représenté par un histogramme d'orientation et de magnitude du Gradient proposée dans [Low04]. C'est un vecteur à 128-dimensions.

FIG. 1.16 – De gauche à droite : Quelques points d'intérêt détectés par le détecteur de Kadir et Brady, le détecteur Harris-Laplacien, le détecteur Harris-Laplacien-Affine.

2. *Apprentissage et sélection de classifieurs* : Chaque objet est représenté par l'ensemble de régions d'intérêt appartenant à l'objet. Tous les vecteurs de caractéristiques déterminés à partir des exemples d'images d'apprentissage d'une classe d'objets sont les points dans un espace de 128

dimensions. Ce nuage de points est modélisé comme un mélange de Gaussiennes qui sont déterminés par un algorithme EM. Chaque point est ensuite assigné à la Gaussienne la plus proche. Un score représentant de combien un cluster est éloigné du fond est déterminé. Les clusters avec les scores les plus grands sont les clusters les plus significatifs et sont considérés comme les clusters des composantes représentatives de la classe d'objets.

3. *Détection et reconnaissance d'objets* : La détection ou la reconnaissance d'objets s'effectue en détectant les points caractéristiques et en calculant les vecteurs de caractéristiques. Le nombre de composantes de la classe d'objets apparues dans l'image décide de la présence ou de l'absence de cette classe d'objets.

7 catégories d'objets différentes ont été utilisées pour l'apprentissage et le test. Cette approche a obtenu un très bon résultat de détection et de reconnaissance. La précision de la reconnaissance est meilleure qu'avec toutes les méthodes présentées dans la littérature, telles que celle de Agarwal et Roth, Hall, Fergus *et al.*, etc. Malgré le fait que l'information sur l'organisation spatiale des composantes ne soit pas considérée, la modélisation par les composantes séparées est très discriminante.

1.2.3 Représentation d'objets basée sur la forme

Les méthodes de représentation d'objet présentées ci-dessus sont basées sur l'apparence. Une alternative est de se baser sur la forme, une caractéristique importante pour décrire des objets. Dans le domaine du traitement et de l'analyse d'images, la forme d'un objet est représentée par une région [KD05] ou un contour [GR96, BM00, BMP02].

Une région peut être décrite par les moments [Hu62, PR92, Wei93] ou la matrice de forme [Gos85]. Un contour peut être décrit par une séquence de points ordonnés [Fre61, PM83], un polygone ou une spline via une approximation polygonale [WL93, BE91, CTCS94] ou spline [IM82] ou un vecteur d'angles entre les tangentes en chaque point du contour [ZR72]. L'appariement des formes consiste à comparer deux descripteurs correspondants. Voir [Lon98, VH99] pour un résumé des méthodes de représentation et de mise en correspondance des formes.

La description de l'objet par une région ou un contour doit supposer que l'objet est segmenté du fond et composé d'une région unique. Ces méthodes sont ainsi limitées aux objets 2D simples. Dans un environnement réel où les objets et le fond sont simultanément complexes et les objets peuvent également subir des changements de point de vue ou de lumière, de telles descriptions ne sont pas suffisamment discriminantes pour reconnaître plusieurs classes différentes d'objets. Nous présentons ci-dessous quelques approches de représentation d'objets basées sur la forme de façon plus sophistiquée, qui sont mieux adaptées aux objets complexes.

Approches statistiques basées sur les points de contour

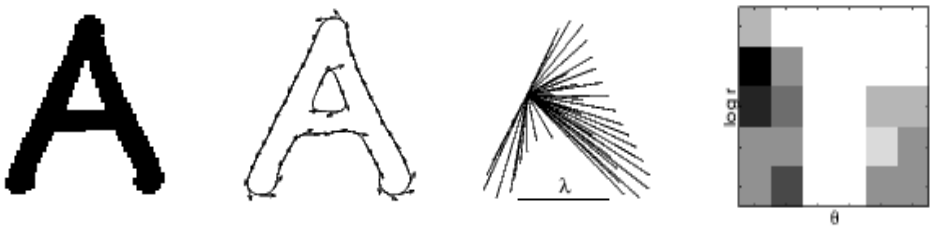
Les approches utilisant les lignes de contour pour modéliser la forme [WL93, BE91, CTCS94, IM82] nécessitent de chaîner des points de contour. Or, la connexité est fortement fragile et sensible au bruit. Certaines approches récentes de représentation de la forme évitent l'étape de chaînage des points de contour en considérant ces points discrets [BM00, TC04].

1. *"Shape context"*[12]

[12] en anglais

La notion de *shape context* est introduite par Belongie *et al.* [BM00]. L'idée de base est la suivante. Supposons que la forme d'un objet soit représentée par un ensemble de n points par exemple des points échantillonnés sur le contour $P = \{p_1, p_2, ..., p_n\}$, $p_i \in R^2$. Si l'on prend un point p_i comme point de référence et l'on construit des vecteurs reliant le point p_i aux points restants, ces vecteurs expriment l'apparence relative de la forme par rapport au point de référence. Lorsque n est suffisamment grand, de tels ensembles de vecteurs représentent exactement la forme entière (voir figure 1.17).

En se basant sur cette idée, Belongie a proposé de décrire chaque point par un shape context qui



(a) (b) (c) (d)

FIG. 1.17 – Calcul du shape context [BM00]. (a) Image binaire d'un objet. (b) Vecteurs de Gradient sur les points de contour. (c) Les vecteurs originaux d'un point aux autres. (d) Le shape context du point considéré.

est un histogramme dans un système de coordonnée *log-polaire*. Pour être invariantes à la position et à l'échelle, toutes les distances sont normalisées par la distance moyenne de toutes les paires de points. La comparaison de deux shape contexts est effectuée selon une technique de comparaison de deux histogrammes.

L'appariement de deux formes s'effectue en appariant les shape contexts des points sur la forme puis estimer une transformation d'une forme à l'autre. Cette technique de mise en correspondance de deux objets basée sur le shape context a été expérimentée sur les formes 2D ainsi que les objets réels dans la base Colombia [MBM01, BMP02]. Les auteurs ont montré une reconnaissance satisfaisante avec un taux d'erreur d'environ 10%.

2. *"OrderType"*[13]

   L'approche de Thureson et Carlsson [TC04] consiste à construire un histogramme d'indices à partir des triplets de points de fort gradient. La procédure se compose de 3 étapes. D'abord, l'image est lissée par un filtre Gaussien, puis son Gradient est seuillé pour obtenir des contours. Ensuite, pour toutes les combinaisons de 3 points de contour, un indice représentant la relation d'angles est calculé (voir la figure 1.18). Finalement, un histogramme d'occurrence de différents indices est construit.

   L'expérimentation de reconnaissance d'objets dans la même base contenant les visages, avions, vélos, voitures utilisées par Fergus *et al.* [FPZ03] montre une performance comparable. Le résultat est significativement amélioré si les images sont segmentées et la modélisation effectuée seulement

[13] en anglais

(a) (b)

FIG. 1.18 – Calcul de OrderType [TC04]. (a) Point sur le contour de l'objet. Chaque order type est calculé sur un triplet de points aléatoires sur le contour. (b) 6 segments crées de 3 points constituent un order type.

sur la région d'intérêt de l'objet.

Remarques sur la représentation par nuage de point : La représentation d'un objet par un nuage de points ne demande pas l'ordre des points. Si la similarité entre deux points est basée sur une mesure locale, la mise en correspondance ne tient pas compte de la cohérence spatiale des composantes de l'objet. Le shape context et l'order type décrivent un point (des points) par rapport aux points qui restent, ainsi la cohérence est prise en compte. Mais celle-ci peut être un inconvénient : elle rend les méthodes sensibles à l'occultation et l'articulation des composantes de l'objet.

Approches structurelles basées sur les caractéristiques topologiques

Les caractéristiques décrivant la topologie des structures sont intéressantes parce qu'elles rendent la représentation plus sémantique. Cette sous-section expose trois approches de représentation structurelles qui influencent fortement notre approche de représentation d'objets.

1. *Approches basées sur le squelette*

   Le squelette est une structure qui permet de représenter à la fois le contour et la région. L'utilisation de squelettes pour représenter la forme a été à l'origine proposée par Blum [Blu67]. L'idée de cette approche est de représenter l'objet sous forme d'un graphe des squelettes en espérant que ses informations les plus importantes soient gardées dans le graphe.

   Une limitation de cette approche est la sensibilité de la détection du squelette par rapport au bruit. Un petit changement dans la forme peut provoquer un grand changement de la topologie du graphe. En outre, Blum a proposé de déterminer les squelettes à partir d'un signal continu. La conversion en espace discret est difficile. Elle peut causer des changements importants par exemple un objet continu est représenté par des graphes discontinus. Ainsi, l'approche de Blum n'est pas applicable en réalité.

   Récemment, Shokoufandeh et ses collègues [SSDZ98] ont amélioré l'approche de Blum pour construire une représentation d'objet sous forme d'arbre comme montré dans la figure 1.19. Dans

cette figure, les points en couleur sont les points sur les squelettes. Chaque couleur correspond à un type de point. Le graphe est construit de telle sorte que les nœuds correspondent à une courbe du squelette. Les axes sont établis entres deux nœuds s'il y a une connexion entre les squelettes. La comparaison de deux objets revient au problème de mise en correspondance des graphes, qui est un problème difficile et très coûteux.

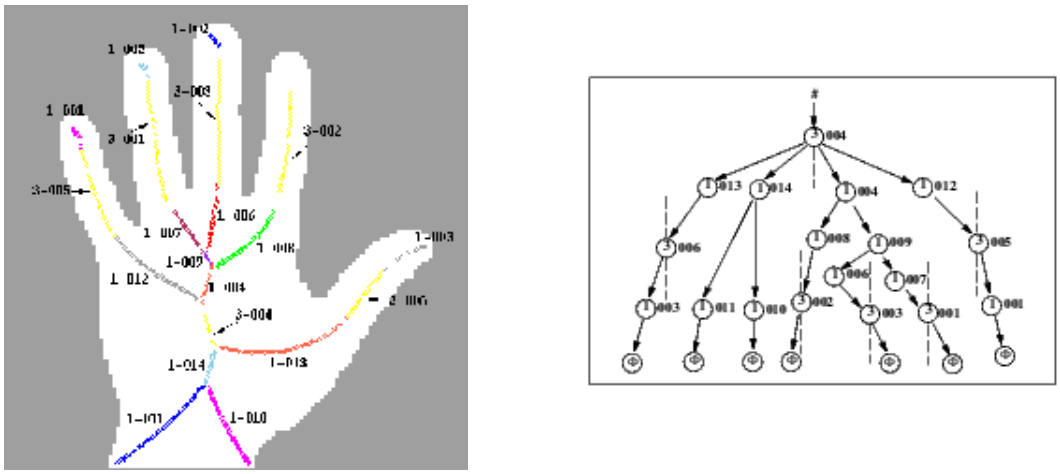


FIG. 1.19 – Représentation d'objet par la décomposition de squelettes à plusieurs échelles [SSDZ98]. La couleur représente le type de squelette. A chaque squelette est assignée une valeur indiquant son caractère significatif.

2. *Représentation hiérarchique à base de points de crête et de pics*

   Motivés par une représentation structurelle de l'objet, dans [CP84], Crowley et Parker ont proposé de représenter un objet par une structure d'arbre construit à partir des points de crête et de pics (voir la figure 1.20).

   Plus concrètement, les éléments pour construire les nœuds de l'arbre sont les caractéristiques telles que les points de crête et les pics détectés à toutes les échelles dans l'espace d'échelle. La connexité des points est étudiée pour construire les liens entre les nœuds. La reconnaissance d'objets consiste à apparier deux arbres.

   Cette approche est intéressante parce qu'en utilisant des caractéristiques détectées à plusieurs niveaux, les caractéristiques au niveau plus haut représentent l'abstraction de la forme, tandis que celles au niveau plus bas représentent les détails. Cette approche est ainsi hiérarchique et permet une représentation grossière-détaillée de l'objet.

3. *Représentation par crêtes et pics à l'échelle optimale*

   Les points de crête et les pics utilisés dans le travail de Crowley et Parker sont détectés à toutes les échelles dans l'espace d'échelle. Dans [LL00], Laptev et Lindeberg ont proposé une approche de représentation d'objets (ie. la main d'une personne) par des crêtes et pics détectés à l'échelle intrinsèque. Chaque pic ou ligne de crête représente une composante de l'objet qui est représentée par une ellipse dont la taille dépend de sa longueur et de son échelle (voir figure 1.21). Les caractéristiques détectées (figure 1.21 à gauche) sont sélectionnées manuellement pour construire le

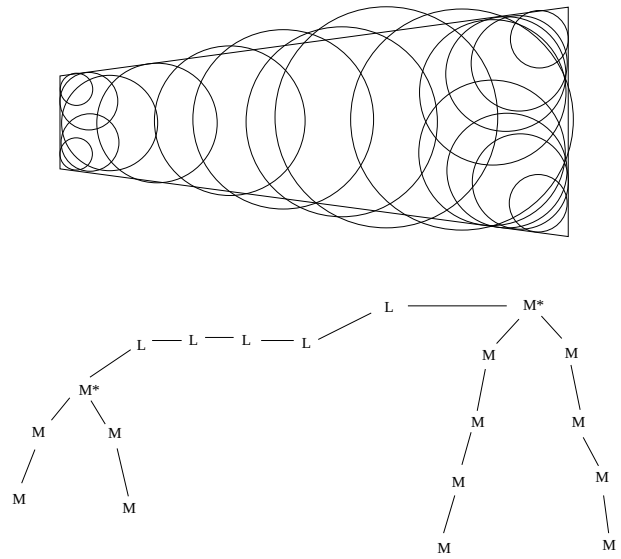


FIG. 1.20 – Une forme rhomboïdale et sa représentation hiérarchique par points de crêtes et pics à quelques échelles [CP84]. L : point de crête. M : pic. M* : pic optimal.

modèle (figure 1.21 à droite).

La main de la personne dans l'application *interaction homme-machine* de Laptev et Lindeberg est un objet structuré. Ainsi les crêtes sont les caractéristiques appropriées pour le représenter. Pourtant, une difficulté dans leur approche est le fait que la détection de points de crête à l'échelle caractéristique peut peut couper accidentellement les lignes de crêtes. La raison en est que l'échelle intrinsèque des points varie le long de la ligne de crête. Si l'espacement entre les échelles n'est pas suffisamment dense (le pas de quantification de l'échelle est grand), il se peut qu'on manque à détecter un point sur la ligne parce que son échelle caractéristique n'est pas dans l'ensemble d'échelles quantifiées.

1.3 Crêtes et la représentation d'objets

Nous venons de voir que la majorité des méthodes de reconnaissance d'objets par apparence ou par forme reposent sur des techniques numériques. Ceci vient de la nature "ponctuelle" des caractéristiques utilisées. La description d'un point par un vecteur de descripteur numériques a donné une reconnaissance très fiable d'objets. Cette description est bien appropriée à la reconnaissance précise d'objets concrets. Pourtant, l'information sémantique de l'objet reste implicite.

Les caractéristiques de type ligne telles que les squelettes, les contours ou les crêtes s'adaptent mieux à une représentation structurelle. L'approche de Crowley et Parker utilise les points de crêtes et les pics, ainsi l'arbre de représentation devient compliqué quand l'objet est complexe [Cro81]. La représentation d'objet par des lignes de crêtes détectées à l'échelle caractéristique proposée par Laptev et Lindeberg [LL00] évite la rédondance des caractéristique. La description de l'objet est plus compacte et l'information de la structure devient plus explicite.

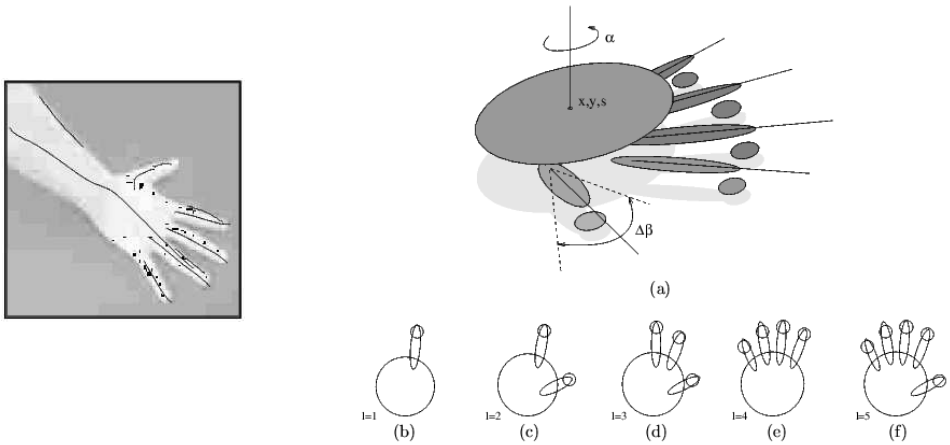


FIG. 1.21 – Représentation d'une main par des crêtes détectées à l'échelle optimale

L'utilisation de crête permet de décrire structurellement des objets et donne la motivation pour étudier des crêtes de façon plus approfondie afin de les appliquer à la représentation. Nous cherchons à explorer les rôles que peut jouer une crête - "ligne d'intérêt naturelle" dans la représentation et comment obtenir une représentation générique, sémantique et structurelle de l'objet à partir des crêtes. Les chapitres suivants vont présenter notre recherche sur la crête et son application à la représentation des objets structurés comme les êtres-humains et les textes.

Chapitre 2

Détection de crêtes et de vallées

La plupart des recherches sur les crêtes sont menées dans le contexte d'applications spécifiques : analyse d'images de vaisseaux sanguins [Pen99, SAN+04], imagerie du crâne [EGM+94, Ebe94, Ebe96, JBAM96, LLSV99, Pen99], recherche de routes dans des images aériennes [MBF92, MAM95], recherche des reseaux de racines de plantes [Roo]. Il y a très peu de travaux sur la représentation d'objets à base de crêtes. Des exemples intéressants sont les travaux de Crowley et Parker pour la modélisation de formes simples [CP84], et Laptev et Lindeberg pour la modélisation des configurations d'une main [LL00].

Dans les travaux de nature plus théorique, on trouve des définitions pour les crêtes dans l'image brute [KvD84, Har83, BPK98], ou dans l'image lissée à une échelle fixe [JBAM96, Pen99, OBS04, SAN+04, SVVV00], à multi-échelle [PSL95], ou à l'échelle caractéristique [CTS95, LL00]. Notons que la détection à une seule échelle est appropriée seulement pour la représentation de structures de même taille. Pour modéliser des objets complexes constitués de plusieurs composantes de taille différente, les crêtes doivent être étudiées dans l'espace d'échelle.

Ce chapitre présente la détection des crêtes à plusieurs échelles dans l'image, en vue de représenter des objets pour la reconnaissance. Nous commençons avec un rappel de quelques notions de géométrie différentielle des surfaces, notamment des types de points selon les courbures principales. Un point de crête est défini sur une surface comme un extremum de la fonction de hauteur, ou de la plus grande courbure principale, dans la direction de cette courbure principale.

Pour détecter des crêtes dans l'espace d'échelle, nous proposons d'utiliser l'opérateur "Laplacien du Gaussien" déjà utilisé pour la recherche de contours. Cette opérateur est paramétrable par l'échelle, et se prête donc facilement à la détection des crêtes correspondant aux structures de taille différente. De plus, il est invariant à de nombreuses transformations : translation, rotation, changement d'échelle. La détection d'un point de crête à une échelle donnée détermine à la fois la position d'une structure et sa taille dans l'image.

L'organisation de ce chapitre est la suivante. La section 1 donne les notions et notations de base de la géométrie différentielle de la surface , et de la représentation de l'image dans l'espace d'échelle. Les sections 2 et 3 donnent un état de l'art des définitions et des techniques de détection de crêtes. La section 4 introduit notre définition des crêtes dans l'espace d'échelle. Nous étudions le comportement du Laplacien en fonction de l'échelle sur un point de crête afin de trouver une bonne échelle pour la détection. La détection de crêtes de cette manière produit des "fausses crêtes" qu'il faut éliminer. Nous

proposons deux techniques pour l'élimination des fausses crêtes : une basée sur le passage par zéro du Laplacien et une autre basée sur le fenêtrage du Laplacien dans la section 5. La section 6 évalue des mesures de crêtes selon les critères, notamment, la détection, la localisation et la continuité des crêtes détectées. Nous effectuons également l'expérimentation de détection des crêtes aux deux cas : à plusieurs échelles et à l'échelle caractéristique. Cette expérimentation a pour but de montrer quels cas les crêtes donnent la meilleure représentation de l'objet.

2.1 Rappel et Notations

Cette section rappelle quelques connaissances de base dans deux domaines différents : la géométrie différentielle et l'espace d'échelle. La construction des expressions différentielles qui capturent les propriétés intrinsèques de l'objet est le sujet de la géométrie différentielle. L'espace d'échelle fournit un cadre de travail utile pour étudier les structures à leur échelle appropriée. La géométrie différentielle de la surface sera étudiée dans l'espace d'échelle pour permettre de représenter les structures de différente taille dans l'image.

2.1.1 Géométrie différentielle

Dérivées partielles

Soit donnée une fonction à 2 variables $L : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$. Les dérivées partielles d'ordre 1 selon x et y de cette fonction sont notées par :

$$L_x = \frac{\partial L(x,y)}{\partial x}, L_y = \frac{\partial L(x,y)}{\partial y}$$

Le vecteur Gradient en un point (x, y) est :

$$\nabla L = \begin{bmatrix} L_x \\ L_y \end{bmatrix}$$

dont $|\nabla L| = \sqrt{L_x^2 + L_y^2}$ et $\alpha = arctan\frac{L_y}{L_x}$ sont respectivement la magnitude et l'angle au point (x, y).

La première dérivée de la fonction L selon une direction v peut être définie par :

$$L_v = \nabla L \cdot \frac{v}{\|v\|}$$

La matrice Hessienne s'écrit :

$$\nabla\nabla L = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix} \tag{2.1}$$

où L_{xx}, L_{yy}, L_{xy} sont les dérivées d'ordre 2 :

$$L_{xx} = \frac{\partial^2 L(x,y)}{\partial x^2}, L_{yy} = \frac{\partial^2 L(x,y)}{\partial y^2}, L_{xy} = \frac{\partial^2 L(x,y)}{\partial xy}$$

La deuxième dérivée de L selon les direction v et w est définie par :

$$L_{vw} = \frac{v}{\|v\|} \cdot \nabla\nabla L \cdot \frac{w}{\|w\|}$$

Surface, courbures principales, directions principales

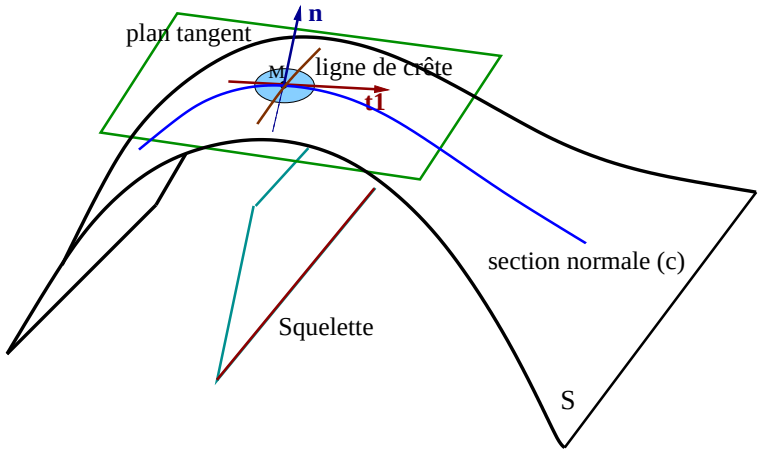


FIG. 2.1 – Géométrie différentielle de la surface

Considérons une surface S étudiée au voisinage d'un de ses points M_0 et définie par une équation de la forme $z = L(x, y)$ où L est (au moins) deux fois continûment dérivable en x et en y (voir figure 2.1).

Si l'on choisit comme repère orthonormé mobile, un repère d'origine M_0 contenant le plan tangent à S en M_0 avec (M_0L) porté par n, vecteur unitaire normal en M_0 à la surface, on peut écrire le développement de Taylor à l'ordre 2 de la côte $z = L(x, y)$ de tout $M(x, y, z)$ infiniment près de M_0 :

$$\begin{aligned} z = L(x,y) &= L(0,0) + [x\frac{\partial}{\partial x}L(0,0) + y\frac{\partial}{\partial y}L(0,0)] \\ &+ \frac{1}{2!}[x^2\frac{\partial^2}{\partial x^2}L(0,0) + y^2\frac{\partial^2}{\partial y^2}L(0,0) + 2xy\frac{\partial^2}{\partial xy}L(0,0)] + ... \end{aligned} \tag{2.2}$$

et vu le choix de notre repère, les dérivées partielles du 1er ordre sont nulles en M_0. D'où, à l'ordre 2 :

$$z = \frac{1}{2!}[x^2\frac{\partial^2}{\partial x^2}L(0,0) + y^2\frac{\partial^2}{\partial y^2}L(0,0) + 2xy\frac{\partial^2}{\partial xy}L(0,0)] \tag{2.3}$$

L'équation (2.3) est une quadrique de la matrice Hessienne $\nabla\nabla L$ (c.f 2.1). Cette matrice a des propriétés suivantes :

- C'est une matrice symétrique, donc diagonisable en base orthonormée.
- Les deux valeurs propres λ_1, λ_2 correspondent aux deux courbures principales de la surface locale. Elles sont déterminées par les expressions suivantes :

$$\begin{cases} \lambda_1 = \frac{L_{xx}+L_{yy}}{2} - \sqrt{\frac{(L_{xx}-L_{yy})^2+4L_{xy}^2}{4}} \\ \lambda_2 = \frac{L_{xx}+L_{yy}}{2} + \sqrt{\frac{(L_{xx}-L_{yy})^2+4L_{xy}^2}{4}} \end{cases} \tag{2.4}$$

D'où on déduit deux vecteurs propres :

$$t_1 = \begin{bmatrix} x_{\lambda_1} \\ y_{\lambda_1} \end{bmatrix} = \begin{bmatrix} \frac{L_{xy}}{\lambda_1 - L_{xx}} \\ 1 \end{bmatrix} \text{ et } t_2 \perp t_1 \tag{2.5}$$

Notons que les vecteurs propres t_1, t_2 se trouvent dans le plan tangent, c'est à dire dans le plan Mxy du repère mobile et local. Si l'on choisit t_1, t_2 comme axe des abcisses du repère, l'équation de la quadrique (2.3) devient :

$$z = \frac{1}{2}(\lambda_1 x^2 + \lambda_2 y^2) \tag{2.6}$$

Les relations de signe et d'ordre de grandeur des valeurs λ_1 et λ_2 indiquent le type de la surface (voir la figure 2.2). Si l'une des deux courbures est nulle, la surface est parabolique. Si les deux courbures sont finies et sont de même signe, la surface est elliptique. Si les deux courbures sont finies et sont de signe opposée, la surface est hyperbolique. Notre approche se base sur cette relation pour distinguer des points de crête des pics.

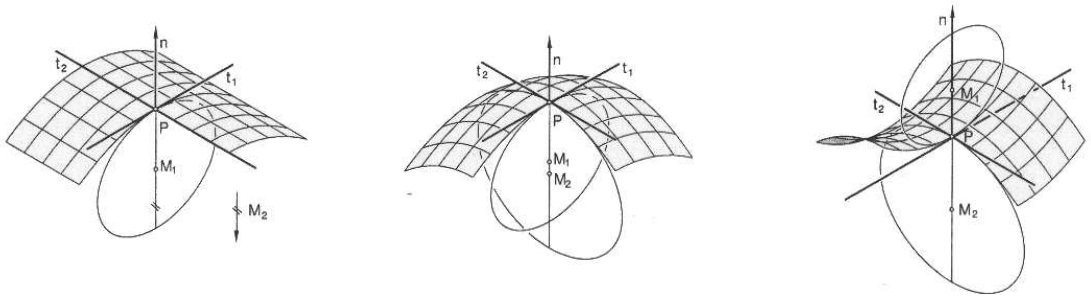


FIG. 2.2 – Classification de la surface via la relation de deux courbures principales. (a) $\lambda_1\lambda_2 = 0$ Parabolique. (b) $\lambda_1\lambda_2 > 0$ Elliptique. (c) $\lambda_1\lambda_2 < 0$ Hyperbolique. Les deux directions principales sont t_1, t_2. Les deux courbures principales sont inverses du rayon des cercles osculateurs.

Courbe de niveau, courbure de la courbe de niveau

La fonction $L(x, y)$ définit une surface $S = \{(x, y) \in \Omega, L(x, y)\}$. Une courbe de niveau l de la surface S est un ensemble des points $S_l = \{(x, y) \in \Omega, L(x, y) = l\}$. La courbure de la courbe de niveau de la surface s'écrit :

$$\kappa(x, y) = \frac{L_y^2 L_{xx} - 2L_x L_y L_{xy} + L_x^2 L_{yy}}{(L_x^2 + L_y^2)^{\frac{3}{2}}} \tag{2.7}$$

Notons que la direction tangente v à la courbe de niveau en un point est perpendiculaire à la direction du vecteur Gradient, c-à-d $v = (L_y, -L_x)^t$.

2.1.2 Représentation multi-échelles de l'image

Nous avons montré au chapitre 2 que les caractéristiques visuelles, de par leur taille dans l'image, possède une échelle intrinsèque. Si on veut modéliser une forme complexe comportant différentes caractéristiques, cette notion d'échelle est cruciale. L'analyse d'une image au contenu inconnu doit donc s'effectuer dans l'espace d'échelle.

L'espace d'échelle a été proposé par Witkin [Wit83], et Koenderink [KvD84]. Le livre de Lindeberg [Lin94] est une très bonne référence, avec des développements mathématiques et algorithmiques sophistiqués.

Koenderink a montré que l'espace d'échelle doit satisfaire l'équation de diffusion de la chaleur qui admet comme unique solution un filtre Gaussien. L'unicité de cette solution a également été montrée, avec des formulations différentes, par Babaurd [BWBD86], par Florack [FtHRKV92] et par Lindeberg [Lin94]. Ces résultats conduisent à l'utilisation du filtre Gaussien pour construire la représentation multi-échelle. Plus précisément, l'image à l'échelle σ, notée $L(x, y; \sigma)$ est le résultat d'une convolution de l'image originale $I(x, y)$ avec la fonction Gaussienne 2D à l'échelle σ $G(x, y; \sigma)$. Dans la suite, on l'appelera le Gaussien de l'image à l'échelle σ :

$$L(x, y; \sigma) = G(x, y; \sigma) * I(x, y)$$

où

$$G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Dans nos expérimentations, nous avons utilisé une "représentation multi-échelle" de l'image qui est un ensemble d'images avec différentes valeurs discrètes d'échelle. La figure 2.3 montre l'exemple d'une image à différentes échelles dans cet espace. On constate qu'en augmentant l'échelle, l'image devient floue, certaines structures disparaissent, d'autres deviennent évidentes.

Comme nous travaillons dans un espace discret, la construction de l'espace d'échelle nécessite de choisir le nombre de niveaux d'échelle et l'espacement entre deux échelles. En principe, l'échelle maximale doit correspondre à la taille maximale des structures dans l'image. Dans le cas d'extrême, cette taille peut égaler à la taille de l'image. Dans un cas où on a des connaissances sur la taille maximale des structures recherchées, on peut limiter l'échelle maximale.

L'espacement des échelles peut être uniforme [MS01] ou exponentiel [CP84]. Crowley a montré dans sa thèse que l'espacement exponentiel avec un facteur $\sqrt{2}$ est un choix raisonnable. Ce choix permet aussi une implémentation efficace de la représentation multi-échelle de l'image [CR03].

Pour ces raisons nous choisissons un espacement exponentiel de $\sqrt{2}$ et un nombre de niveaux d'échelle $K = \log_2(M\text{x}N)$ où M et N sont les dimensions d'image. Les échelles sont donc 1, $\sqrt{2}$, 2, $2\sqrt{2}$, ..., $\sqrt{2}^{K-1}$.

2.2 Formes de crête

D'un point de vue mathématique, la notion de crête peut être définie de différentes manières, qui sont plus ou moins adaptées à la recherche de caractéristiques visuelles dans une image multi-échelle. En général, un point de crête est défini comme un extremum local dans une certaine direction. Une ligne de crête est un ensemble de points de crête connexes. *Cette définition de crête, vue en rapport avec celle*

FIG. 2.3 – Représentation multiéchelle de l'image. De haut en bas, de gauche à droite : l'image originale et les images lissées aux échelles $\sigma_1 = 1, \sigma_2 = 2, \sigma_3 = 2\sqrt{2}, \sigma_4 = 4, \sigma_5 = 4\sqrt{2}$.

de points d'intérêt naturels - extrema dans toutes les directions - nous fait appeler des crêtes - lignes d'intérêt naturelles.

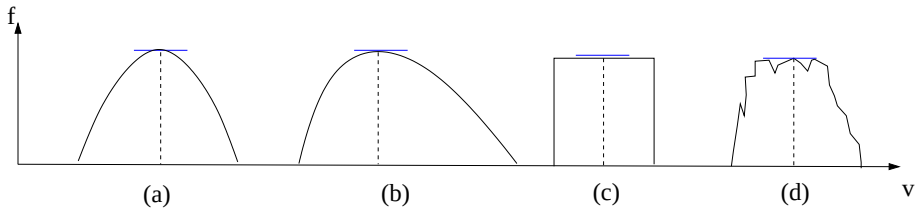
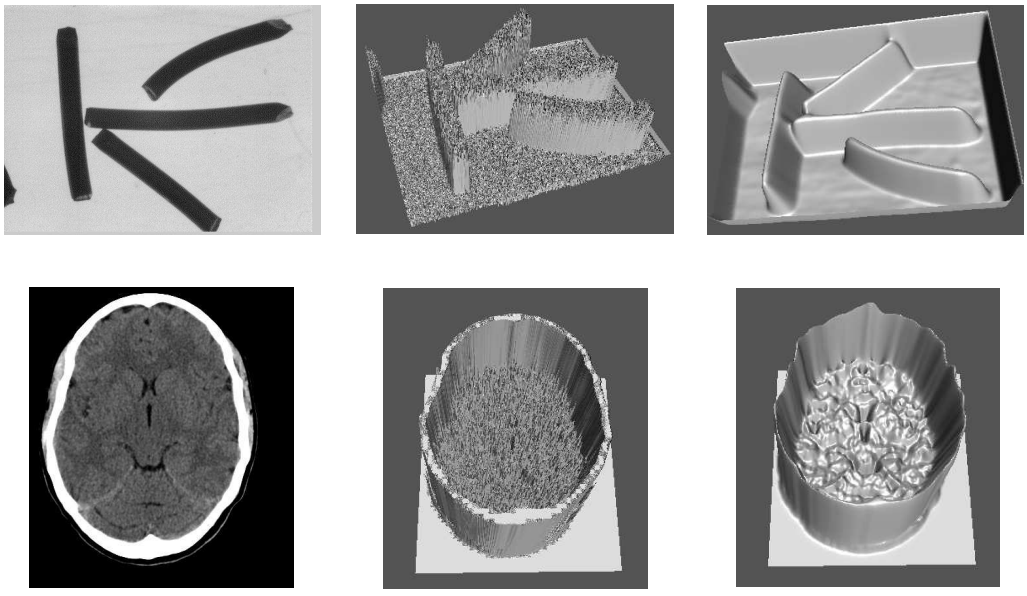


FIG. 2.4 – (a) Modèle idéal de crête. (b, c, d) Crêtes réalistes.

La figure 2.4a montre la section perpendiculaire à la direction de la crête en un point de crête idéal. La surface d'un point de crête réaliste peut avoir des formes variées (figures 2.5b,c,d). Si nous appliquons le critère d'extremum sur ces types de surface, aucun point de crête (figure 2.4c) ou plusieurs points bruités de crête (figure 2.4d) sont trouvés. La raison en est que la surface n'a pas été étudiée à la bonne échelle.



(a) (b) (c)

FIG. 2.5 – (a) Image originale. (b) Représentation 3D de l'intensité de l'image. (c) Représentation 3D de l'image lissée à l'échelle $\sigma = 8$ (en haut) et $\sigma = 4$ (en bas).

Si nous filtrons les surfaces par un filtre Gaussien de taille σ appropriée, les "bruits" dans la figure 2.4d disparaissent et toutes les surfaces ont une forme de la surface d'une crête idéale. La figure 2.5 montre un exemple de surfaces associées à une image qui ne contiennent pas de point de crête. Les crêtes ne ressortent quand on considère la surface à l'échelle appropriée. D'où une remarque importante : *sans précision de l'échelle la notion de crête n'a guère de sens*.

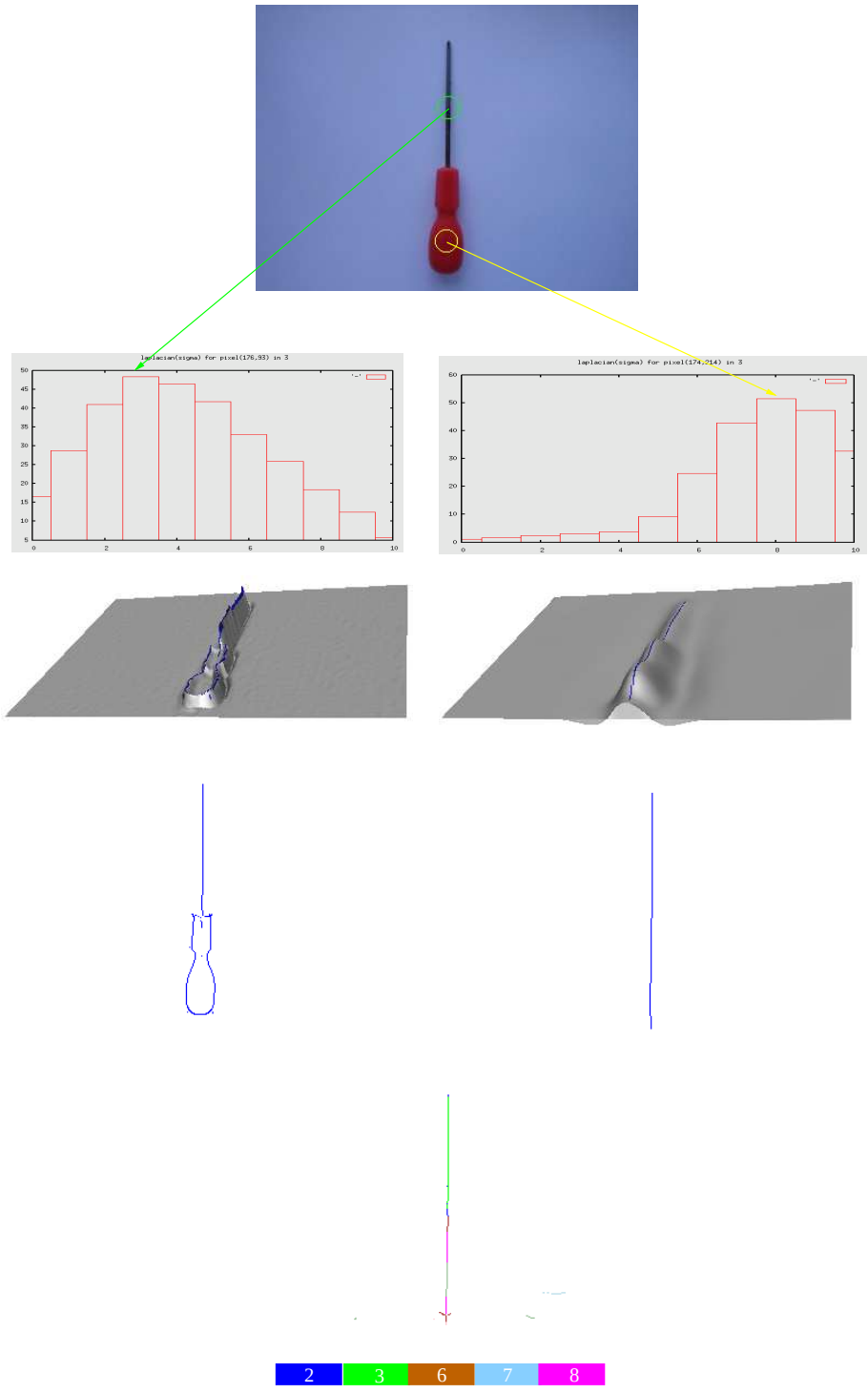


FIG. 2.6 – Première ligne : Image d'un tournevis. Deuxième ligne : les profils de Laplacien du Gaussien selone l'échelle aux points (176,193) - cercle vert et (176, 214) - cercle jaune, on constate un maximum (les niveaux 3 et 8 correspondent aux échelles $2\sqrt{2}$ et 16 respectivement) dans chaque profile correspondant à l'échelle caractéristique du point. Troisième ligne : la représentation 3D du Laplacien du Gaussien superposé par des crêtes détectées à l'échelle $\sigma_1 = 2\sqrt{2}$ (à gauche) et σ_2 (à droite) (quatrième ligne). Cinquième ligne : la crête détectée à l'échelle caractéristique. Les couleurs représentent le niveau d'échelle optimal en chaque point de crête.

La figure 2.6 montre l'exemple dans lequel on nécessite de détecter des crêtes à certaines échelles appropriées pour pouvoir capturer des structures de différentes tailles d'un objet. Il s'agit d'un tournevis composé de structures allongées de tailles différentes : la taille de la lame dans l'image est de 6 pixels, l'échelle appropriée pour capter la lame est donc $\sigma_1 = 2\sqrt{2}$; la taille du manche est de 21 pixels, ce qui demande une détection à l'échelle $\sigma_2 = 16$.

La quatrième ligne de la figure 2.6 montre les crêtes détectées à deux échelles $\sigma_1 = 2\sqrt{2}$ et $\sigma_2 = 16$. On remarque toute de suite que le manche ne ressort qu'à grande échelle. A cette échelle, comme la lame est lissée, on trouve également une partie de crête représentant la lame. La cinquième ligne montre la présence des caractéristiques aux échelles caractéristiques. Le changement de couleur le long de la crête principale indique la variation de l'échelle caractéristique des points sur la crête.

2.3 Quelques définitions classiques des crêtes

Bien que le concept de crête a été traité depuis le 19ième siècle dans la littérature mathématique (Saint Venant - 1852 [SV52]), les définitions pour les crêtes dans la littérature en vision par ordinateur prennent des points de départs très variés. Plusieurs termes tels que *ridge, valley [Har83, CTS95, Ebe96], watershed [VS91, Ste96], crest line [SLE93, MAM95, BGT96], ravine [BPK98], crease, seperatrice [LLSV99]* ont été utilisés pour *le concept crête*, avec des approches et des définitions différentes.

Nous présentons dans cette section deux familles de définitions de crêtes : une basée sur la fonction de hauteur qui caractérise plutôt le changement du signal de l'image ; une autre basée sur les courbures principales qui représentent les propriétés intrinsèques structurelles de la surface. Ces définitions faites en 2D sont extensibles à un espace à n dimensions. Notons que toutes les définitions de crêtes présentées ci-dessous s'appliquent à une surface. La notion d'échelle n'a pas encore ou modestement été étudiée dans chaque définition. Ainsi, ces définitions ne s'appliquent qu'à la surface de l'image originale, ou de l'image lissée à une échelle particulière.

2.3.1 Définitions basées sur la fonction de hauteur

Il est naturel de considérer une image 2D comme une carte de niveau avec l'intensité comme paramètre de hauteur. Les points sur une ligne de crête sont les maxima du signal de l'image dans une certaine direction. Pour détecter des points de crête dans l'image, il suffit de détecter les extrema locaux directionnels.

Définition de Saint-Venant

Un des premiers travaux sur l'identification d'une ligne de crête sur une surface, datant depuis 1852, est dû à De Saint Venant [SV52]. Saint Venant identifie un point de crête à l'endroit où la magnitude du Gradient le long d'une ligne de niveau est minimale (figure 2.7). La condition de Saint Venant peut être écrite de façon compacte comme suit :

$$L_{vw} = 0 \text{ et } |L_{ww}| < |L_{vv}| \tag{2.8}$$

où L est la fonction qui définit la surface considérée et w est la direction du gradient et v est la direction perpendiculaire.

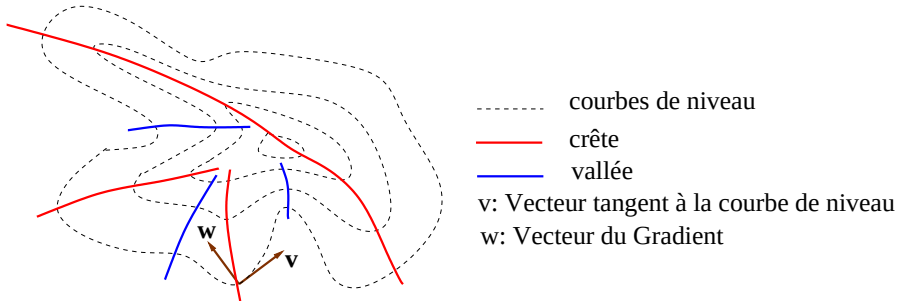


FIG. 2.7 – Courbes de niveau et les crêtes/vallées déterminées selon la condition de Saint-Venant.

Définition de Haralick

La condition de Saint-Venant est plus tard reformulée par Haralick [Har83] pour appliquer à l'analyse d'image. Haralick identifie un point de crête ou de vallée aux endroits où la première dérivée selon la direction qui extremise la deuxième dérivée passe par zéro. Les dérivées sont calculées de façon analytique en approximant le signal d'image par un polynôme cubique.

Un problème irrémédiable dans le travail de Haralick est que l'utilisation des masques pour approximer les cœfficients du polynôme représentant la surface locale peut modifier brutalement des propriétés de la surface. Par exemple, une surface définie par une ligne blanche de largeur d'un pixel sur fond noir est transformée en une surface du type selle. Trouver une bonne approximation apparaît impossible. Ceci est un défaut pertinent de cette approche.

Les travaux les plus proches de celui proposé par Haralick sont dus à Paton [Pat75] et Hsu *et al.* [HMB78]. Paton a approximé la surface de l'image $L(x, y)$ par une fonction quadratique et identifié des points de crête aux endroits où $L(x, y)$ a une valeur significative dans une seule direction. Comme l'approximation selon Paton est quadratique, la définition de crête ne peut être appliquée qu'au centre des pixels. L'approximation de la surface par une cubique comme fait Haralick permet en revanche d'avoir une précision sous-pixellique.

Variantes du modèle d'Haralick

Une variante de l'idée d'Haralick est proposée dans [CTS95]. Chinveeraphan *et al.* localisent les points de crête au lieu où la direction du Gradient coïncide avec une des directions principales de la matrice Hessienne. Une autre variation se trouve dans [JBAM96]. En constatant que le vecteur perpendiculaire au Gradient pointe vers la direction de plus grande courbure, Maintz *et al.* ont proposée d'extraire des extrema locaux de la dérivée d'ordre 2 selon la direction perpendiculaire au vecteur de Gradient. Cette dérivée est calculée selon la formule suivante :

$$\begin{aligned} L_{vv} &= \frac{1}{\|v\|^2}(v.\nabla)^2 L \\ &= -\frac{L_y^2 L_{xx} - 2L_x L_y L_{xy} + L_x^2 Lyy}{L_x^2 + L_y^2} \end{aligned} \tag{2.9}$$

où $v = (L_y, -L_x)^t$ le vecteur perpendiculaire au vecteur Gradient.

Dans le but de segmentation d'image (surtout les images médicales), Kalitzin *et al.* [KStHRV01], Staal *et al.* [SKA+02], Stoeckel *et al.* [SVVV00] ont proposées de classifier les points sur la surface de l'image en classes "crête", "vallées", "selle"[1]. L'identification des ces points se base sur la recherche des extrema de l'intensité ou d'un canal de couleur de l'image selon la direction des vecteurs propres de la matrice Hessienne.

Extensions de définition d'Haralick

L'idée de Haralick a été étendue en [MPL93] et surtout en [EGM+94, Ebe94, Ebe96]. Eberly *et al.* extraient des crêtes de r dimensions dans l'image d-dimensionnelle ($1 \leq r \leq d$). Un point de crête de r-dimensions se trouve au lieu où la fonction $L(x_1, x_2, ..., x_d)$ admet un extremum local dans r directions prises de r vecteurs propres de la matrice Hessienne correspondante. Cette caractérisation est appelée sous le terme en Anglais *height condition* et reformulée comme la suivante : Supposons que $|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_d|$ les valeurs propres de la matrice Hessienne $\nabla\nabla L$, et $v_1, v_2, ..., v_d$ les vecteurs propres correspondants, un point de crête rD est caractérisé par :

$$\forall i \in \mathcal{I}_{d-r}, \nabla L.v_i = 0$$

et

$$\begin{cases} \lambda_i < 0 \text{ si crête} \\ \lambda_i > 0 \text{ si vallée} \end{cases} \tag{2.10}$$

où $\mathcal{I}_j$ l'ensemble des indices entiers de 1 à j.

Une contribution intéressante de Eberly *et al.* dans [EGM+94] est l'utilisation des crêtes détectées à plusieurs résolutions d'image pour faire la segmentation d'image 3D (MR-images). Cette approche d'analyse des crêtes est remarquable en deux points : (1) l'introduction de l'échelle en étudiant des crêtes qui permet d'étudier le changement topologique des structures ; (2) l'analyse des relations de recouvrement des régions caractérisées par des crêtes via échelles.

2.3.2 Définitions basées sur la courbure

La définition de crête basée sur la courbure repose sur l'analyse des propriétés géométriques de la courbe dans un plan et de la surface dans R^3. L'idée est de choisir des points tels que la surface au voisinage soit de courbure maximale selon une direction.

Définition de Gauch

Dans [GP93], Gauch *et al.* utilisent les outils de la géométrie différentielle pour étudier la surface image. Ils identifient des points de crêtes comme extrema locaux de courbure de sa courbe de niveay de la surface[2]. Ces points sont alors connectés d'un niveau à un autre niveau et constituent une ligne de vertices appelée ligne de crête ou ligne de vallée. La courbure d'une courbe de niveau dans l'image est calculée selon (cf. 2.7).

[1] saddle point. Il s'agit de point en quel deux courbures principales de la surface associée sont finis et sont de signe opposé.
[2] Parfois appellé "vertex condition" en anglais [SLL00]

La recherche de l'extremum de la courbure κ consiste à chercher le passage par zéro de la dérivée de la courbure selon la direction tangente v à la courbe de niveau, $v = (L_y, -L_x)^t$.

$$D_v\kappa = \frac{\nabla\kappa \cdot v}{\|v\|}$$

Cette expression peut être développée en combinaison non-linéaire de dérivées d'ordre 1, 2, 3 de l'image originale. Dans [GP93], ces dérivées sont calculées de façon analytique en approximant la fonction d'image par une spline cubique.

Nous constatons que l'approche de Gauche *et al.* est similaire à l'approche proposée par Saint-Venant. En effet, les points de magnitude du Gradient minimale sont ceux détecté par Gauche. En outre, nous remarquons que l'expression (cf. 2.7) est juste une normalisation de (cf. 2.9) par la magnitude du Gradient.

Variantes de la définition de Gauch

Des définitions similaires à celle de Gauch *et al.* peuvent se trouver dans les travaux de Monga *et al.* [MBF92, MAM95], Bruce *et al.* [BGT96], Lang *et al.* [LBBK97], Belyaev *et al.* [BPK98], Miller *et al.* [MF99], Maintz *et al.* [JBAM96], etc.

Dans [MBF92, MAM95], un point de crête est identifié au lieu de la plus grande courbure principale le long de sa direction. Les courbures principales et les directions principales sont calculées par les filtres récursifs. Ce travail est pour but d'extraire des réseaux de lignes fines dans une image satellite (ie. routes) ou une image médicale (ie. vaisseaux de sang). Avec ces types d'images, les lignes sont assez fines et de même largeur. Ainsi, les crêtes extraites à une seule échelle sont suffisantes pour représenter ces lignes (dans l'article $\sigma = 1$).

Dans [BPK98], Belyaev *et al.* ont développé une formule pour calculer des points de crête sur une surface implicite. Une ligne de crête est une courbe d'intersection entre deux surfaces implicites : celle originale et celle des extrêma locaux des courbures principales le long leurs directions principales. Cette approche de calcul des crêtes se développe d'une manière très mathématique, ce qui est approprié à la caractérisation des surfaces définies par des formules mathématiques. L'application en analyse d'images demande d'approximer le signal image par une fonction. Or l'approximation permettant de conserver les propriétés géométriques de la surface n'est pas toujours évidente (problème d'approximation).

Extension de définition de crête en d-dimensions par Eberly *et al.*

L'extension de la définition de crête en d-dimensions basée sur la courbure est similaire à celle basée sur la fonction de hauteur que nous avons présentée précédemment. Pour l'image en d-dimensions, il est nécessaire de généraliser la courbe de niveau en hypersurface. Un hypersurface de niveau l consiste en ensemble de points $S_l = \{x \in \Omega : L(x) = l\}$. Alors, si $|\kappa_1| \geq |\kappa_2| \geq ... \geq |\kappa_d|$ sont les courbures principales de l'hypersurface S_l et $t_1, t_2, ..., t_d$ les directions principales correspondantes, un point de crête de r-dimensions est caractérisé par :

$$\forall i \in \mathcal{I}_{d-r}, \nabla\kappa_i \cdot t_i = 0$$

$$\begin{cases} \text{si } t_i^t \cdot \nabla\nabla\kappa_i \cdot t_i < 0 \text{ et } \kappa_i > 0 \text{ point de crête} \\ \text{si } t_i^t \cdot \nabla\nabla\kappa_i \cdot t_i > 0 \text{ et } \kappa_i < 0 \text{ point de vallée} \end{cases} \tag{2.11}$$

2.3.3 Bilan sur les définitions de crête existantes

Le tableau 2.1 résume les méthodes de détection de crête dans la littérature. Deux critères permettent de classer les méthodes : la mesure de crête[3] et la direction selon laquelle la mesure de crête admet un extremum local. Nous faisons les remarques suivantes :

Méthode	Critères d'identification	
	Mesure de crête	Direction
Saint-Venant (1852)	Magnitude du Gradient	Courbe de niveau
Haralick (1983)	Dérivée directionnelle d'ordre 1	Extremum de la dérivée directionnelle d'ordre 2
Eberly *et al.* (1996)	Intensité	Vecteurs propres de la matrice Hessienne
Maintz *et al.* (1996)	Dérivée d'ordre 2	Perpendiculaire au vecteur Gradient
Monga *et al.* (1992)	Courbure principale	Direction associée
Gauche *et al.* (1993)	Courbure de la courbe de niveau	Vecteur tangent à la courbe de niveau
Lang *et al.* (1997)	Courbure principale	Direction principale associée
	Courbure principale	Section verticale générée par la direction principale
Belyaev *et al.* (1998)	Courbure principale	Direction principale associée

TAB. 2.1 – Résumé de méthodes de détection de crêtes

1. **Dualité** : Un point de vallée peut être défini de la même manière qu'un point de crête. La distinction des points dépend du signe de la plus grande valeur propre de la matrice Hessienne : crête si $\lambda_1 < 0$ et vallée si $\lambda_1 > 0$. Il suffit donc d'étudier seulement le concept de crête. Au long de cette thèse, nous allons le plus souvent utiliser le terme "crête" pour désigner à la fois crête et vallée ; parfois, le terme spécifique "vallée" sera utilisé. Notons aussi que ce qui apparaît comme une crête positive dans l'image originale (ou lissée avec un filtre Gaussien) apparaît comme vallée négative dans le Laplacien (voir la figure **??**).
2. **Invariance** : La géométrie différentielle montre que les courbures principales de la surface locale associée à un point sont invariante par des transformations telles que la translation, la rotation, le changement uniforme de l'intensité. Ainsi, les crêtes détectées selon les définitions basées sur la courbure sont invariantes à telles transformations. La définition de crête basée sur la fonction de hauteur n'est pas invariante à la rotation.
3. **Echelle** : La plus ancienne définition de crête (ie. celle de Saint Venant 1852) est purement mathématique et n'aborde pas la notion d'échelle. Celle-ci est apparue lorsqu'on a voulu détecter les structures pour l'analyse d'images. La plupart des travaux fixent une échelle pour la détection de crête car les images qu'ils analysent contiennent des structures de largeur assez uniforme. La détection de crête à l'échelle fixe ne permet pas de capturer des structures de taille variée dans l'image naturelle. Peu de travaux ont analysé des crêtes dans l'espace d'échelle pour construire une représentation à la fois globale et détaillée de l'objet.
4. **Evaluation de performance des mesures** : L'expérimentation pour évaluer des mesures de crêtes a été réalisée sur les surfaces mathématiques définies. Maintz *et al.* [JBAM96] et Eberly [EGM+94,

[3] ridgeness en anglais

Ebe94] ont montré que la définition de crête basée sur la fonction de hauteur donne des crêtes plus "intuitives". L'évaluation quantative des mesures n'a jamais été réalisée. En fait, cette dernière nécessite un corpus des images dans lesquelles les crêtes sont marquées manuellement. Ceci est difficile parce que la perception humaine ne peut pas détecter correctement la position de crêtes, qui ne sont pas identiques avec celles des squelettes.

2.4 Définition de crête basée sur le Laplacien du Gaussien

La section 2.1.2 a montré que dans le contexte de description des structures dans l'image, la notion de crête n'a guère de sens si elle est étudiée dans échelle. La détection de crête dans une image $I(x, y)$ à une échelle σ peut s'effectuer en détectant les extrema de la fonction de hauteur $L(x, y; \sigma) = G(x, y; \sigma) \otimes I(x, y)$ ou la courbure principale de la surface définie par $L(x, y; \sigma)$ selon une certaine direction. Dans cette thèse, nous proposons d'utiliser le Laplacien du Gaussien (voir la figure 3.8 pour la forme de LoG en 2D et 3D). Nous commençons d'abord par étudier le comportement de Laplacien du Gaussien en un point de crête puis la définition algorithmique.

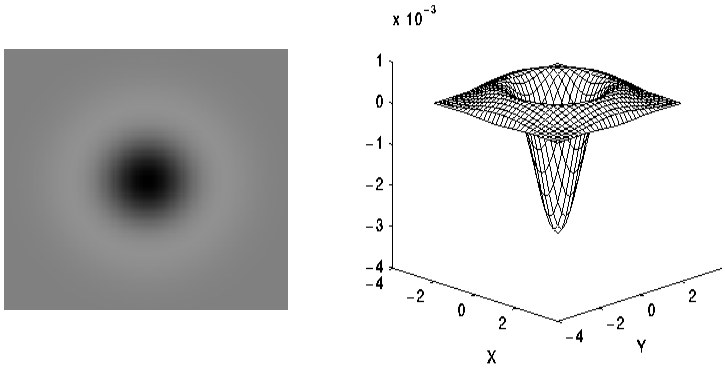


FIG. 2.8 – Laplacien de Gaussien dans 2D et 3D

2.4.1 Comportement de l'opérateur Laplacien du Gaussien

Considérons une image $I(x, y)$ dans l'espace d'échelle construit avec $K = \log_2(\text{NxM})$ niveaux d'échelle espacés exponentiellement, le niveau d'échelle i correspondant à l'échelle $\sigma_i = \sqrt{2}^i$.

La figure 2.9a montre une bande noire sur fond blanc. La taille de l'image est de 257x257 pixels, la largeur de la bande est de 32 pixels, tous les pixels à l'intérieur de la bande ont une intensité égale à 255, l'intensité sur les autres points est nulle. Dans ces conditions, on s'attend à détecter une crête verticale au centre de la bande $x = 128$ à l'échelle $\sigma = 32/2 = 16 = \sqrt{2}^8$ qui correspond au niveau $i = 8$.

A droite de la figure 2.9 sont présentés les profils, à 11 niveaux d'échelle, du filtre Gaussien, le long d'une ligne traversant l'image de gauche à droite, perpendiculairement à la bande . Nous constatons qu'initialement, le point au centre de la bande est de même intensité que les points voisins. Il n'y a aucun extremum. Quand l'échelle augmente, le filtre Gaussien lisse petit à petit la bande et un maxima

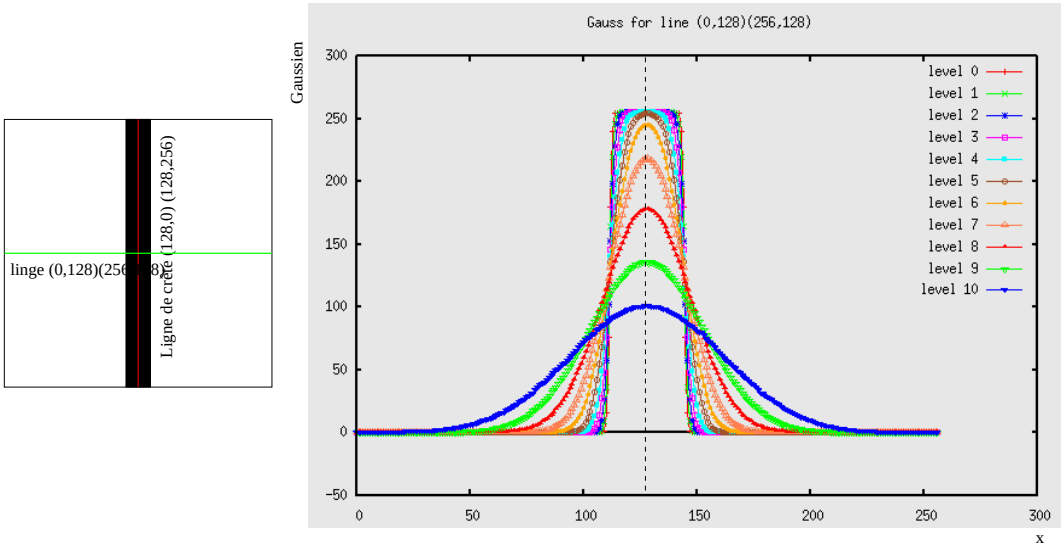


FIG. 2.9 – A gauche : une bande noire sur fond blanc. A droite, les profils des Gaussiens le long de la ligne verte à 11 échelles différentes.

au centre apparaît. Ce maxima correspond à un point de crête. Plus l'échelle est grande, plus la valeur du maximum est petite. Ceci est inverse pour le cas d'une bande blanche sur le fond noir. Le maximum existe à la même position à plusieurs niveaux d'échelles.

Regardons maintenant le comportement de la réponse de l'opérateur Laplacien du Gaussien en un point sur la ligne $x = 128$ (figure 2.10). Un extremum apparaît quand l'échelle augmente. Quand l'échelle est égale à la moitié de la largeur de la bande ($\sigma = 32/2 = 16$ correspondant au niveau 8), il y a deux passages par zéro de Laplacien exactement aux points de contour. De plus, la réponse du Laplacien est la plus forte au point central. Cette échelle est l'échelle caractéristique du point considéré.

2.4.2 Définition de crête basée sur le Laplacien du Gaussien

L'étude du comportement du paragraphe précédent montre que le Laplacien du Gaussien est capable de détecter les changements directionnels, simultanément de localiser la largeur de la structure correspondante. Nous proposons donc une définition d'un point de crête basée sur le Laplacien du Gaussian :

Définition d'un point de crête : *Soit donnée une image $I(x,y)$. A une échelle σ le lissage de l'image définit une surface $\{x, y, L(x,y,\sigma\}$ où $L(x,y;\sigma) = G(x,y;\sigma) \otimes I(x,y)$. Un point (x,y) est un point de crête à l 'échelle σ si le **Laplacien du Gaussian** de l'image à l'échelle σ, $\nabla^2 L(x,y;\sigma)$ admet un extremum local dans la direction correspondante à la plus grande courbure λ_1 de la surface associée.*

$$\begin{cases} \textit{si } \lambda_1 < 0 \textit{ point de crête} \\ \textit{si } \lambda_1 > 0 \textit{ point de vallée} \end{cases} \tag{2.12}$$

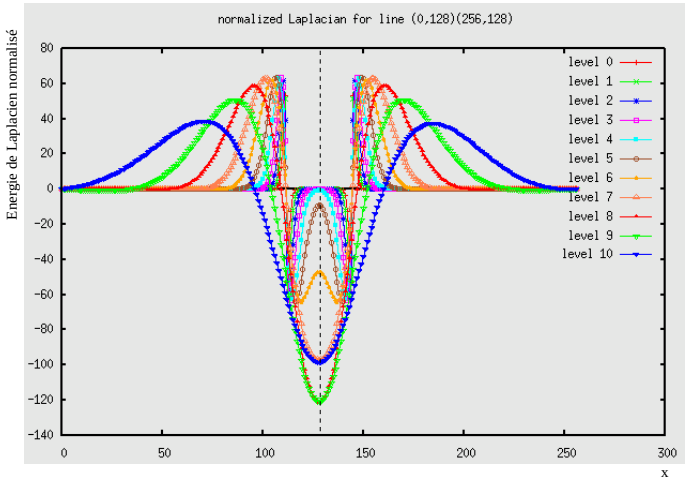


FIG. 2.10 – Les profils de Laplacien de Gaussien le long de la ligne verte à 11 échelles. Chaque niveau d'échelle i correspond à une valeur $\sqrt{2}^i$.

Dans la partie gauche de la figure 2.11, on montre une image de fils de fer et la détection à l'échelle $\sigma = 16$ par l'opérateur Laplacien du Gaussien. Les vallées (les lignes bleues) détectées représentent parfaitement les troncs de fils de fer dans différentes directions. La largeur des fils dans l'image est de 45 pixels. L'échelle la plus proche de la moitié de la taille de fils est 16. Dans l'image de directions, les différentes directions de courbures principales sont représentées par des couleurs différentes. Le long du fil en haut à droite et celui en bas à droite de l'image, la direction change à cause de la discrétisation de l'espace de direction (voir la section 2.5.3). Le chaînage de points de crête doit tenir compte de cette situation pour ne pas manquer la ligne de crête longue.

Dans la partie droite de la figure 2.11 on montre une image d'un objet plus sophistiqué : un vélo, l'image de direction principale t_1 en chaque point (chaque couleur représente une direction), le Laplacien de l'image à l'échelle $\sigma = 4\sqrt{2}$ et les crêtes et vallées détectées à cette échelle (les lignes bleues et les lignes rouges). Nous constatons que les crêtes trouvées représentent des structures significatives de l'objet. Le critère (2.12) produit des crêtes "fines" et bien continues.

Dans un cas où les structures sont arrondies, le Laplacien du Gaussien donne une réponse maximale dans toutes les directions. Nous distinguons ce cas en seuillant le rapport des valeurs propres $|\frac{\lambda_1}{\lambda_2}|$ selon la classification de type de surface étudiée dans la section **??**. Si $|\frac{\lambda_1}{\lambda_2}| < seuil$ alors pic, sinon crête. Dans l'exemple d'une fleur (figure 2.12), nous avons seuillé le rapport des valeurs propres par un seuil 2.0. Les crêtes "arrondies" deviennent les pics et les vallées deviennent les trous.

2.5 Elimination de fausses crêtes

On appelle fausse crête toutes les crêtes qui ne correspondent pas à une structure dans l'image. Par exemple, certaines lignes rouges dans la figure **??** sont des fausses crêtes.

Nous voulons qu'une vraie crête représente une structure physique de l'objet. Dans l'image, à cause

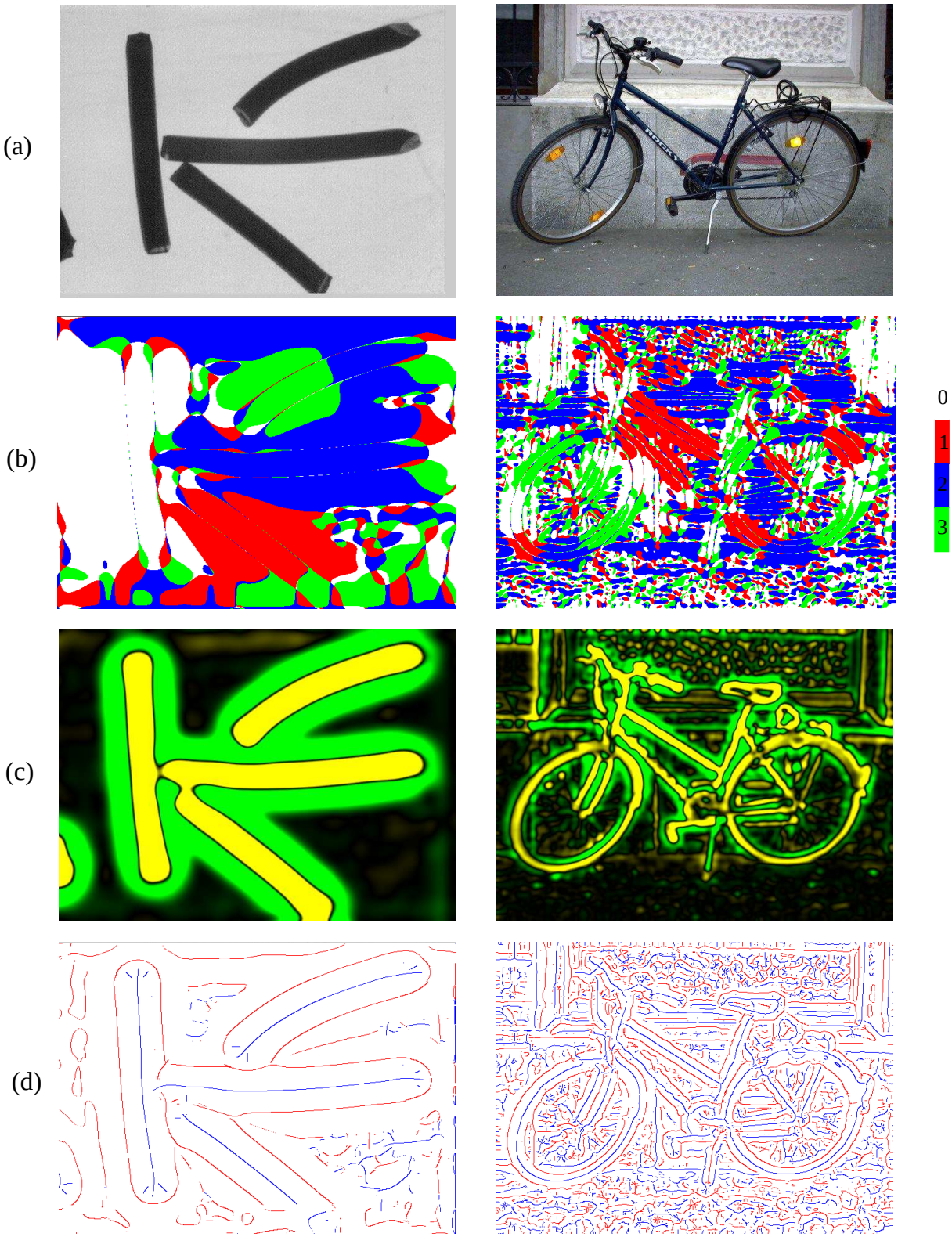


FIG. 2.11 – (a) Images originales. (b) Images de directions. (c) Images de Laplacien du Gaussien (jaune : valeur positive, vert : valeur négative). (d) Images de crêtes (rouges) et vallées (bleues). La détection est réalisée à l'échelle $\sigma = 16$ (à gauche) et $\sigma = 4\sqrt{2}$ (à droite).

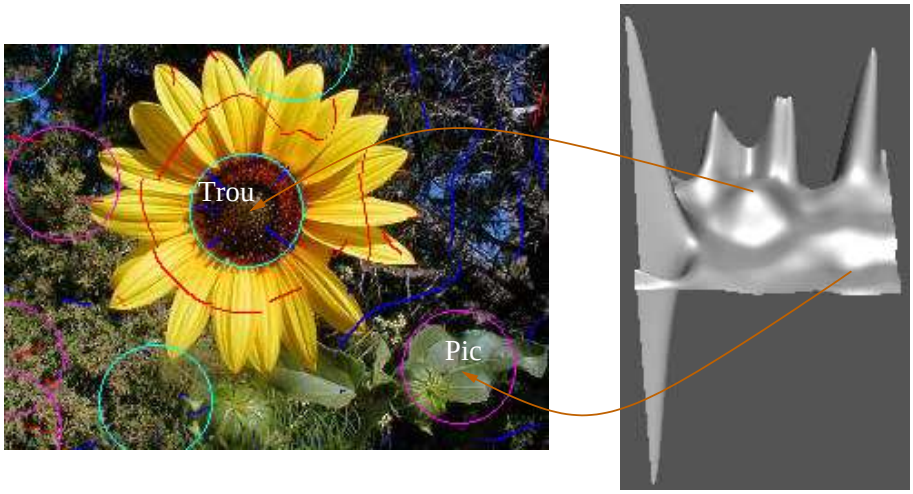


FIG. 2.12 – (a) Image d'une fleur superposée par des crêtes (lignes rouges), vallées (lignes bleus), pics (cercle violet), trous (cercle cyan) à l'échelle $\sigma = 16\sqrt{2}$. (b) Représentation 3D de Laplacien à l'échelle $\sigma = 16\sqrt{2}$. Tous les points dont le ratio de deux valeurs propres est inférieur à 2.0 ne sont pas considérés comme point de crête.

de l'arrangement entre des objets et de l'effet de luminosité, une vraie crête peut ne correspondre à aucune structure (le vide entre deux structures physiques, l'ombre). Nous discutons dans la suite sur les fausses crêtes qui sont provoquées par la méthode de calcul.

Considérons une Gaussienne $f(x,y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$. Pour tous les points sur la ligne $x = 0$, les deux directions principales sont alignées aux deux axes x et y. La courbure dans la direction x est la plus grande. Elle est égale à la deuxième dérivée de f selon x :

$$\lambda_1 = f_{xx} = \frac{1}{\sqrt{2\pi}\sigma^3}(\frac{x^2}{\sigma^2} - 1)e^{-\frac{x^2}{2\sigma^2}}$$

λ_1 est maximale selon la direction x si la première dérivée passe par zéro et la deuxième dérivée est négative :

$$\lambda_{1x} = f_{xxx} = 0 \text{ et } \lambda_{1xx} = f_{xxxx} < 0$$

Nous avons :

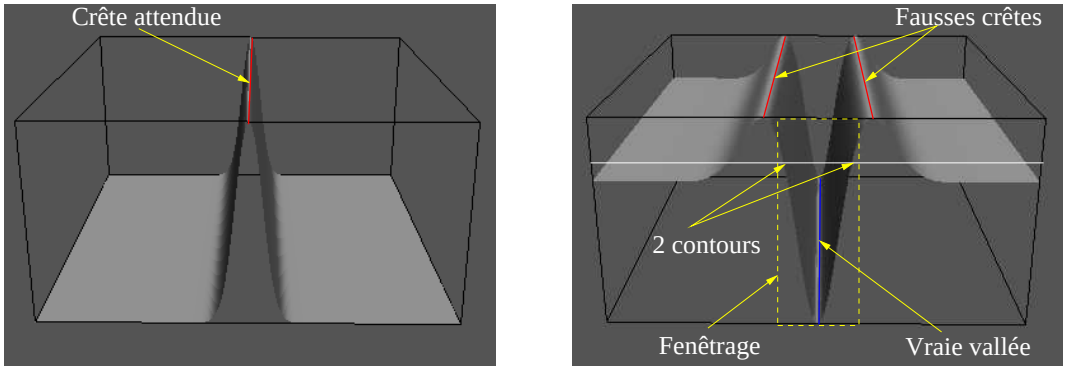
$$f_{xxx} = \frac{1}{\sqrt{2\pi}\sigma^3}(\frac{3x}{\sigma^2} - \frac{x^3}{\sigma^4})e^{-\frac{x^2}{2\sigma^2}}$$

qui a trois solutions :

$$x = 0, x = \sqrt{3}\sigma, x = -\sqrt{3}\sigma.$$

D'ailleur,

$$f_{xxxx} = \frac{1}{\sqrt{2\pi}\sigma^3}(\frac{x^4}{\sigma^6} - \frac{6x^2}{\sigma^4} + \frac{3}{\sigma^2})e^{-\frac{x^2}{2\sigma^2}}$$



(a) (b)

FIG. 2.13 – (a) Crête dans la Gaussien $f(x,y) = e^{-x^2/2\sigma^2}/2\pi\sigma^2$ où $\sigma = 16$. (b) Laplacien de Gaussien $\nabla^2 G$. A gauche, la ligne rouge $x = 0$ est une vraie crête, qui apparait à droite comme vraie vallée (en bleu). Les deux lignes rouges $x = \pm\sqrt{3}\sigma$ sont deux fausses crêtes.

nous avons $f_{xxxx}|_{x=0} = \frac{3}{\sigma^2} > 0$, $f_{xxxx}|_{x=\pm\sqrt{3}\sigma} = \frac{-6}{\sigma^2}e^{-\frac{3}{2}} < 0$. Donc, $x = 0$ est une ligne de crête selon la définition basée sur la fonction de hauteur, et est une ligne de vallée selon la définition basée sur la courbure et le Laplacien. Mais, en $x = \pm\sqrt{3}\sigma$, le Laplacien du Gaussien ainsi que la courbure présentent 2 extrema locaux qui produisent deux crêtes dans les sens opposés. Ces crêtes ne correspondent à aucune structure attendue (voir la figure 2.11).

Pour représenter un objet par des crêtes, les fausses crêtes ne sont pas dangereuses. En fait, si un objet modèle est représenté par des vraies crêtes et des fausses crêtes, un objet réel l'est aussi. Ainsi, malgré que l'information soit redondante, elle n'influence pas sur le résultat de mise en correspondance. Pourtant, pour avoir une représentation compacte et propre, l'élimination de fausses crêtes est souhaitable.

Notons que la définition de crête basées sur la courbure principale produit également les fausses crêtes car la deuxième dérivée directionnelle agit avec le même comportement que le Laplacien. Pourtant, dans les articles, les auteurs n'ont pas montré les crêtes et les vallées en même temps. Ce problème n'est donc pas posé comme vrai problème de recherche. Nous proposons deux méthodes pour éliminer des fausses crêtes.

2.5.1 Méthode 1 : Passages par zéro

Normalement, une crête "intuitive" représente une structure dans une scène. Une structure physique est limitée dans le plan d'image par deux contours physiques. La structure "vide" au milieu de deux structures physiques est limitée par deux contours de chaque structure. La distinction de la crête physique et crête "artificielle" peut être basée sur cette propriété. En tous cas, une vraie crête doit se trouver au milieu de deux contours.

De cette observation, nous proposons un critère qui permet de distinguer vrais et faux points de crête. Le critère est le suivant :

Un point de crête (x,y) est vrai si l'extremum de Laplacien du Gaussien se trouve au milieu de ses deux passages par zéro dans la direction de la courbure principale de la surface en (x,y). Les passages

par zéros doivent se trouver à la distance limitée $\sigma + \epsilon$ depuis ce point. ϵ est de petite valeur.

2.5.2 Méthode 2 : Fenêtrage de Laplacien de Gaussien

En observant le profil de Laplacien du Gaussien, nous remarquons que les fausses crêtes correspondent à deux minima juste à côté de deux points de contour. Pour éviter ces crêtes, il faudrait rendre nulles les réponses près du contour à l'extérieur de la structure. Nous proposons de fenêtrer le Laplacien du Gaussien par un rectangle de taille $[-\sqrt{2}\sigma, \sqrt{2}\sigma]$ (voir la figure 2.13).

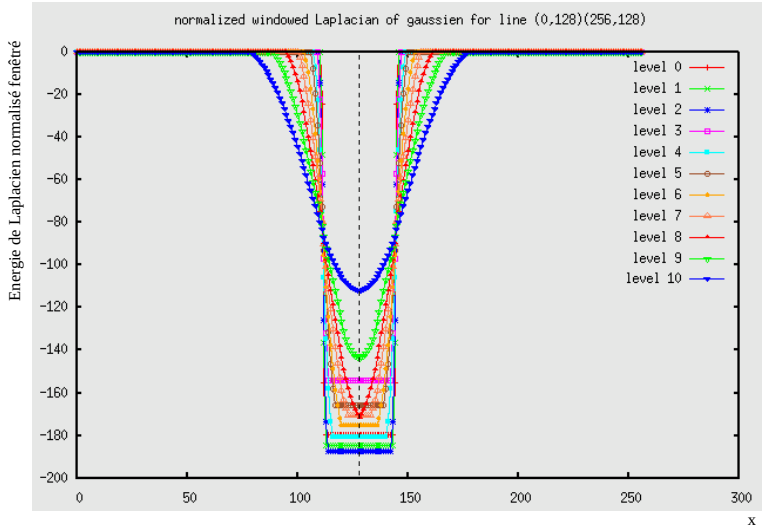


FIG. 2.14 – Laplacien du Gaussien fenêtré et normalisé le long de la ligne verte dans la figure 2.9.

Une propriété intéressante de Laplacien du Gaussien fenêtré par rapport au Laplacien est que l'extremum local n'apparaît qu'au moment où l'échelle est égale à la moitié de la taille de structure, pas aux échelles avant ni aux échelles après (voir la figure 2.14). Alors, l'échelle à laquelle apparaît la première fois un extremum local de Laplacien du Gaussien est l'échelle caractéristique. L'utilisation du Laplacien fenêtré évite de détecter des crêtes représentant une même structure mais existant à différentes échelles. L'algorithme de détection de crête en utilisant le Laplacien fenêtré ressemble à celui basé sur le Laplacien.

La figure 2.15 compare le résultat d'élimination des fausses crêtes par deux méthodes proposées sur deux images : l'image de fils de fer à gauche et l'image d'un vélo à droite. Nous constatons que la méthode 2 élimine des fausses crêtes de façon plus raisonnable que la méthoede 1. En fait, la méthode 2 ne "touche" pas les vraies crêtes. Elle réduit seulement les réponses en fausses crêtes, ce qui sont ensuite enlevées par un seuillage. La méthode 1 enlève des fausses crêtes près du contour du cadre du vélo. Mais des "vraies" crêtes sont éliminées aussi. On s'aperçoit ici que le critère de présence de deux passages par zéro dans deux directions opposées apparaît trop strict et rend discontinues certaines vraies crêtes (par exemple le cadre dessous, deux roues du vélo).

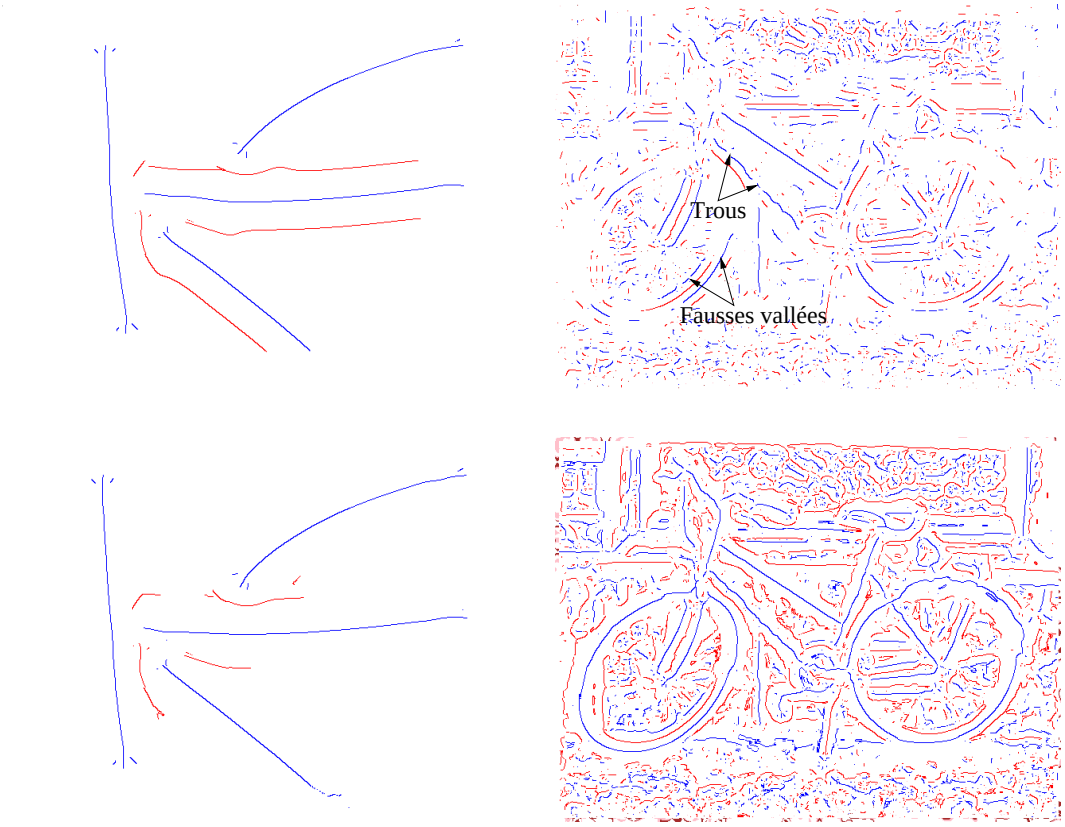


FIG. 2.15 – Crêtes et vallées détectées par le Laplacien du Gaussien fenêtré à l'échelle $\sigma = 16$ avec un seuil du Laplacien 35.0, dans l'image de la figure **??**. (a) Par la méthode 1 : le Laplacien du Gaussien et la vérification des passage par zéros. (b) par la méthode 2 : le Laplacien de Gaussien fenêtré. Notons que des fausses crêtes subsistent encore dans la figure (a). Dans la figure (b), les crêtes dans le vide entre deux fils de fer sont correctes, il s'agit des crêtes blanches sur fond noir.

2.5.3 Algorithme de détection des points de crête

Calculer des dérivées

Les dérivées des images lissés sont calculées de façon très efficace en utilisant le filtre récursif proposé par Vliet *et al.* [VYV98]. Le calcul des courbures principales et des directions principales est direct par les formules (2.4) et (2.5).

Notons que dans les formules (2.4) et (2.5), les dérivées sont déterminées dans le repère local et mobile, pas dans le repère d'image. En réalité, nous supposons que dans un voisinage infiniment près d'un point considéré, le plan tangent est approximativement parallèle au plan d'image. Dans ce cas, toutes les dérivées d'ordre 2 dans le repère local et mobile restent les mêmes qu'au repère d'image. L'expérimentation nous montre que cette hypothèse donne des résultats satisfaisants de détection de crête en image 2D.

Discrétiser les directions

Pour simplifier le calcul des extrema, l'espace des directions est discrétisé en 4 comme montré dans la figure 2.16. Nous nous intéressons seulement à l'orientation pas à la direction. Ainsi, une direction x sera codée en une des 4 valeurs $\{0, 1, 2, 3\}$.

Avec la discrétisation de l'espace des directions, la précision du calcul des extrema est limitée. Le calcul des points voisins sur une ligne de pente $tan(\alpha)$ est très approximatif.

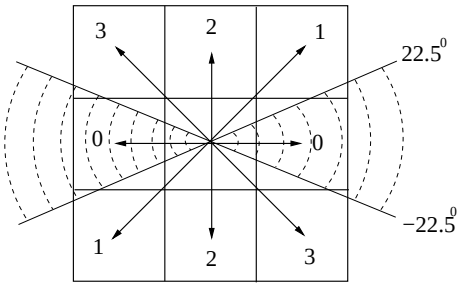


FIG. 2.16 – L'espace de direction est discrétisé en 4.

.

Déterminer des extrema locaux

Une mesure de crête (la fonction de hauteur, la courbure de la courbe de niveau, la courbure principale, le Laplacien, etc) admet un extremum local en (x, y) dans une direction d si elle est plus grande ou plus petite que celle en d'autres points voisins de (x, y) selon la direction d. Comme l'espace de direction est dicrétisé, cette étape n'est pas difficile.

Etapes de l'algorithme

L'algorithme de détection est implémenté de façon générique afin de permettre de tester plusieurs mesures de crête ou de combiner des critères par un opérateur booléen. Il se compose de trois étapes principales :
Entrée : Image de niveau de gris ou d'un canal de couleur

1. Calcul des dérivées : L_x, L_y,L_{xx}, L_{yy}, L_{xy}, $\nabla^2 L$, λ, les mesures de crête à toutes les échelles.
2. Détermination des directions principales de la matrice Hessienne t_1, t_2.
3. Recherche des extrema de la mesure de crête selon les directions principales t_1.

Sortie : Image binaire dont le pixel (x, y) vaut 0,1,2,3,4 montre la présence d'un point du fond, de crête, vallée, pic, trou respectivement.

Les applications de crêtes présentées dans les chapitres 4 et 5 utilisent les lignes de crêtes. A partir de l'image binaire des points, nous réalisons une analyse de connexité de points de même type pour former les lignes.

2.6 Expérimentation

Cette section consiste à évaluer la performance de

1. Quatre mesures de crêtes : le Gaussien de l'image, la courbure de la courbe de niveau, la courbure principale, le Laplacien du Gaussien. La question posée est quelle est la meilleure mesure pour la détection des crêtes ?
2. La représentation des structures par crêtes à l'échelle optimale ou à multi-échelle. Le problème se pose avec les crêtes détectées à une ou plusieurs échelles. Comment représenter les objets à partir de ces crêtes ? Les crêtes détectées à l'échelle caractéristique sont-elles meilleures pour la représentation de l'objet que l'ensemble des crêtes détectées pour l'objet à différentes échelles ?

2.6.1 Evaluation quantitative des mesures de crêtes

Critères d'évaluation

Les mesures de crêtes sont évaluées sur les surfaces définies mathématiquement. Par contre, aucune évaluation quantitative n'a été proposée sur les surfaces correspondantes aux images réelles. La raison en est qu'il est difficile de créer un corpus des images avec des crêtes étiquetées à la main.

Malgré cette difficulté, on veut qu'à la sortie de l'algorithme de détection, les crêtes doivent satisfaire certain nombre de propriétés telles que :

- *Invariance* aux rotations, aux translations et aux changements uniformes de la lumière.
- *Bonne détection*. Il détecte des crêtes qu'on attend et il ne détecte pas de fausses crêtes.
- *Bonne continuité*. Les crêtes continues devraient être obtenues à partir de structures continues.

L'invariance de détecteurs de crête peut être démontrée théoriquement. Les valeurs propres sont invariantes aux transformations citées.

Pour évaluer la détection et la continuité des crêtes, nous choisissons une base d'images de test contenant des textes. La raison de ce choix est qu'un texte contient plusieurs caractères qui apparaissent comme des traits longs, distingués du fond. Ces traits peuvent être représentés par des crêtes détectées à l'échelle reliée à la largeur du trait. L'utilisation des images de textes permet de déterminer faicilement des crêtes attendues ainsi que leur continuité.

La détection est évaluée en comparant le nombre de crêtes détectées et le nombre de traits calculés manuellement dans le texte. La continuité est mesurée par le nombre de trous dans un trait continu de texte. Ces mesures sont définies par les formules suivantes :

$$\text{Rappel(Détection)} = \frac{\sum_{i=1}^{N} \text{nbtruedetect}_i}{\sum_{i=1}^{N} \text{nbridge}_i}$$

$$\text{Continuité} = \sum_{i=1}^{N} \text{nbtrou}_i$$

Où, $N = 50$, le nombre d'images de texte dans la base.

Notons que nous calculons seulement le rappel de la détection. La précision est difficile à calculer parce que si le texte se trouve sur un fond complexe, il n'est pas évident de déterminer les vraies crêtes

des fausses crêtes. Pour certaine application telle que la détection de texte, le rappel est plus important que la précision car on veut détecter toutes les crêtes correspondant aux traits de caractères. De cette manière, le nombre de trous ne se calcule que sur les crêtes correspondant aux traits des caractères, pas sur les autres crêtes. Il est le nombre minimal des points nécessaires pour remplir les trous dans les crêtes.

Rappel de 4 mesures de crêtes utilisées pour la comparaison

Quatre mesures de crêtes sont utilisées pour la comparaison. Elles sont : le Gaussien de l'image, la courbure de la courbe de niveau κ, la courbure principale λ et le Laplacien Lap. Les trois dernière mesures sont calculées sur le Gaussien de l'image originale.

- **Gaussien de l'image**
$$L(x, y; \sigma) = G(x, y; \sigma) \otimes I(x, y)$$
- **Courbure de la courbe de niveau**
$$\kappa(x, y) = -\frac{L_y^2 L_{xx} - 2L_x L_y L_{xy} + L_x^2 L_{yy}}{(L_x^2 + L_y^2)^{\frac{3}{2}}}$$
- **Courbure principale**
$$\lambda_1 = max\{\frac{L_{xx} + L_{yy}}{2} \pm \sqrt{\frac{(L_{xx} - L_{yy})^2 + 4L_{xy}^2}{4}}\}$$
- **Laplacien**
$$Lap = L_{xx} + L_{yy}$$

Pour détecter les crêtes à une échelle σ, les mesures ci-dessus sont calculées à l'échelle σ et normalisées en fonction de l'échelle σ. Concrètement, une dérivée d'ordre n est multipliée par σ^n. Dans tous les cas, la direction principale est la direction correspondante à la plus grande valeur propre de la matrice Hessienne.

Résultat de l'évaluation sur la base d'images de textes

Nous avons détecté des crêtes dans 50 images naturelles. Il s'agit d'images de journaux télévisés et d'une séquence de course de voiture. Chaque image contient des textes ou groupes de textes comme illustre la figure 2.17. Dans ces images, la largeur des traits des caractères varie de 4 pixels à 16 pixels. Ainsi, nous détectons des crêtes dans un intervalle d'échelle $[2, 8]$. Il se peut qu'une image contienne plusieurs textes de tailles différentes. Dans ce cas, nous détectons les crêtes à toutes les échelles qui correspondent à la moitié des tailles des traits présentes dans l'image. Ces tailles sont déterminées manuellement.

Le tableau 2.2 montre le résultat obtenu en utilisant les 4 mesures de crêtes ci-dessus. Il faut noter que nous n'avons pas seuillé l'énergie du Laplacien dans ces tests. Il y a 384 crêtes que l'on souhaite détecter à partir de 50 images. Toutes les mesures donnent un bon rappel d'ordre 98%. Le rappel n'est pas de 100% pour les deux raisons suivantes : Premièrement, pour chaque image nous avons fixé l'échelle ou un ensemble d'échelles pour la détection. Ainsi, certains traits trop minces (ie. ceux qui ont la largeur

plus petite que deux fois l'échelle) sont manqués (voir par exemple dans le texte Marlboro, le premier trait à gauche du M est trop mince). Deuxièmement, certains traits trop flous ou de même couleur que le fond sont difficiles à détecter.

Parmi quatre mesures testées, la courbure de la courbe de niveau donne le plus de réponses bruitées, non significatives près des crêtes principales. Les crêtes détectées par cette mesure se localisent loin du milieu des traits. Elles sont coupées à même les structures à l'intérieur qu'elles trouvent assez uniformes en couleur (voir figure 2.18). La courbure principale et le Laplacien donnent une réponse similaire. En fait, la plupart des traits des caractères sont horizontaux ou verticaux. Ainsi, le Laplacien est égal à la dérivée d'ordre 2 selon x ou y, qui est égale à la courbure principale. Dans cette base d'images, le Laplacien détecte des crêtes légèrement plus continue que la courbure principale (74 trous contre 91 trous).

FIG. 2.17 – Les images de texte extraites dans la base de 50 images.

La figure 2.18 montre deux exemples de texte, un net sur le fond simple, facile à détecter et un autre flou, subi une projection perspective, difficile à détecter. En comparant les crêtes détectées à l'intérieur des caractères, nous trouvons que la courbure de la courbe de niveau détecte beaucoup d'artefacts. Le La-

Mesure	#Vraies Crêtes/#Crêtes réelles	#Trous
Gaussien de l'image	377/384(0.982)	104
Courbure de Courbe de niveau	373/384(0.971)	200
Courbure Principale	379/384(0.987)	91
Laplacien	381/384(0.992)	74

TAB. 2.2 – Comparaison des mesures selons les critères de détection et de continuité des crêtes.

placien détecte des crêtes un peu plus significatives que la courbure principale (par exemple le caractère R dans TELEDIARIO et FOSTER'S).

2.6.2 Comparaison sur d'autres types d'image

Nous réalisons la comparaison de performance de 3 mesures de crête : κ, λ_1, Lap. Le Gaussien de l'image définit la surface sur laquelle les crêtes sont détectées. Toutes les mesures sont calculées à une échelle fixe prédéfinie. Le but est de montrer la performance de détection de crête par différentes mesures et l'expressivité des crêtes pour la représentation des structures de l'objet.

Images médicales Une application de la crête qui a attiré l'intérêt des chercheurs dans le domaine d'imagerie est la mise en correspondance une image CT[4] avec une image MR[5] d'une même personne pour combiner les informations partielles complémentaires. En effet, l'image CT représente précisément les os tandis MR différencie mieux les tissus mous. Comme la tête du patient n'a pas de la même position et orientation dans les deux scanners MR et CT, la mise en correspondance est nécessaire. Pour cela, les auteurs détectent une crête circulaire dans l'image CT et une vallée dans l'image MR. La crête et la vallée sont appariées afin de trouver une transformation géométrique entre deux images.

La première ligne de la figure 2.19 montre une image MR et la représentation 3D du Gaussien de cette image à l'échelle $\sigma = 4$. Intuitivement, nous voulons détecter une ligne de vallée circulaire correspondant au centre de crâne. Idem pour l'image CT (voir figure 2.20).

Les quatre lignes en bas de chaque figure montrent l'image de mesure de crête, le résultat de détection de crête sans seuiller l'énergie de Laplacien, avec le seuillage et la superposition des crêtes sur l'image originale et l'agrandissement d'une partie de l'image pour comparer la performance de chaque mesure de crête utilisée. Trois colonnes présentent le résultat obtenu à partir de trois mesures de crêtes : la courbure de courbe de niveau, la courbure principale et le Laplacien respectivement.

Nous constatons que la courbure de courbe de niveau produit des crêtes non-significatives dans les deux cas CT et MR. Dans le cas de l'image MR, la courbure principale détecte une crête parfaitement continue correspondant au centre de crâne. Le Laplacien a des difficultés à détecter une crête avec la bonne continuité. En fait, dans la région encadrée par le rectangle, le signal de l'image varie légèrement. A l'échelle $\sigma = 4$, la surface de la courbure principale a des caractéristiques de crête tandis qu'à cette échelle la surface de Laplacien est encore très moutonnée. Quand l'échelle est suffisamment grande ($\sigma = 8$), la surface associée au Laplacien a la caractéristique de la surface d'une crête continue. Ceci explique que l'échelle ne soit pas appropriée, le Laplacien ne produit pas les réponses attendues. Ici,

[4]Computed Tomography
[5]Magnetic Resonance

FIG. 2.18 – De haut en bas, l'image originale du texte, le résultat de détection à l'échelle $\sigma = 4$ (avec le texte " TELEDIARIO") et l'échelle $\sigma = 2$ (avec le texte "FOSTER'S") en utilisant 4 mesures : le Gaussien, la courbure de la courbe de niveau, la courbure principale, le Laplacien.

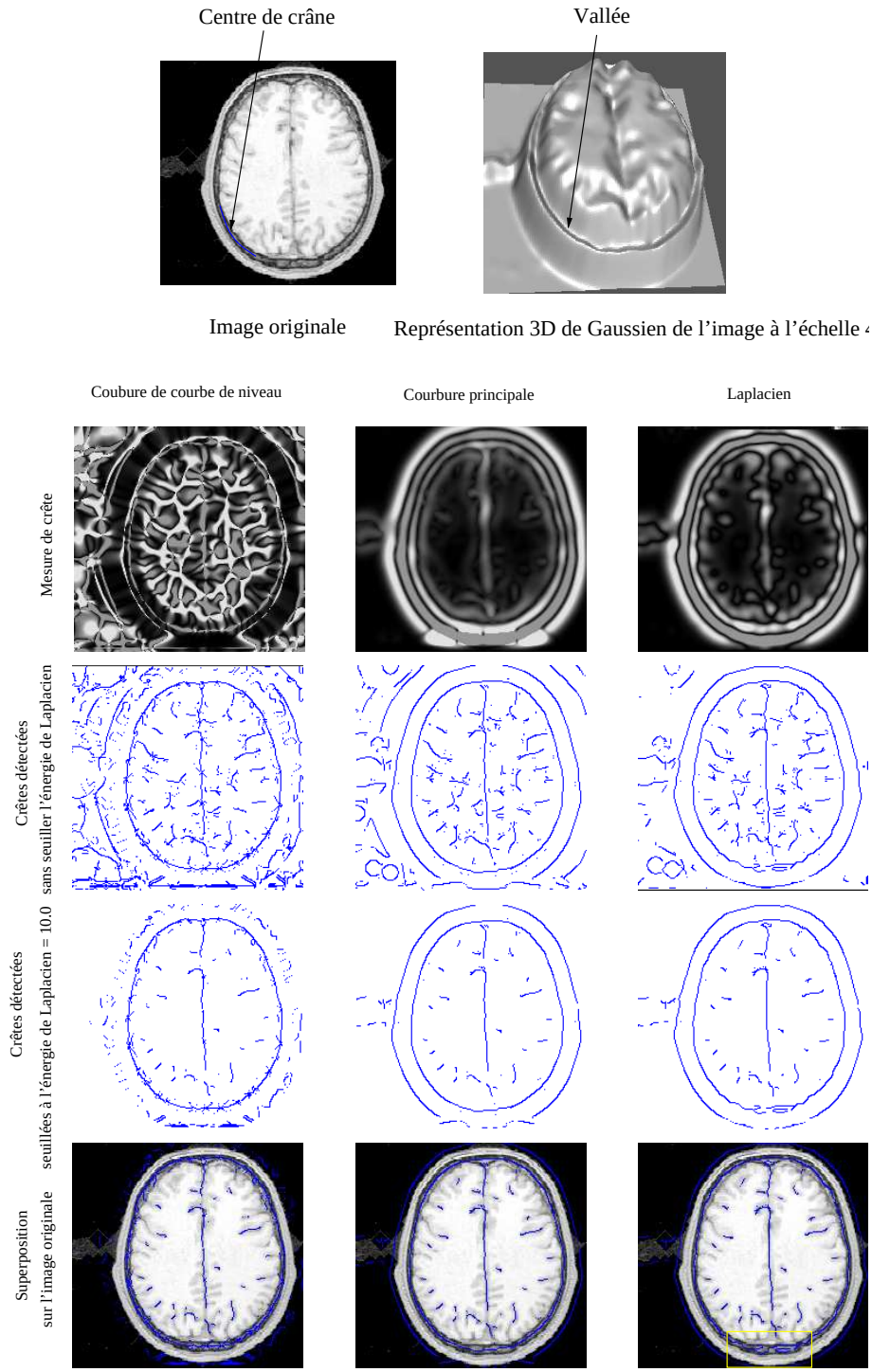


FIG. 2.19 – Détection de l'image MR à l'échelle $\sigma = 4$.

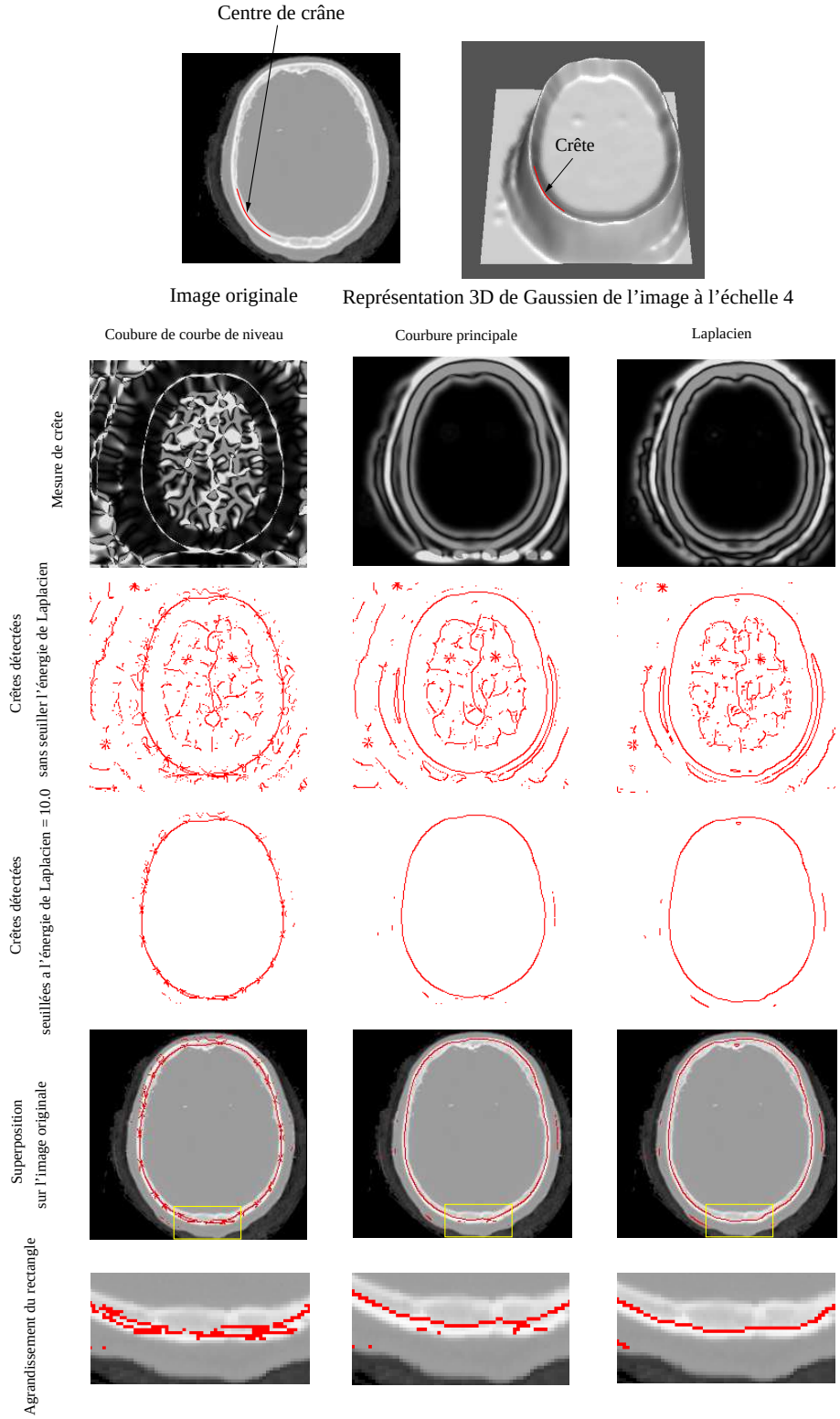


FIG. 2.20 – Détection de l'image CT à l'échelle $\sigma = 4$.

comme la structure n'est pas homogène, l'échelle doit être grande pour que le lissage de la région soit assuré.

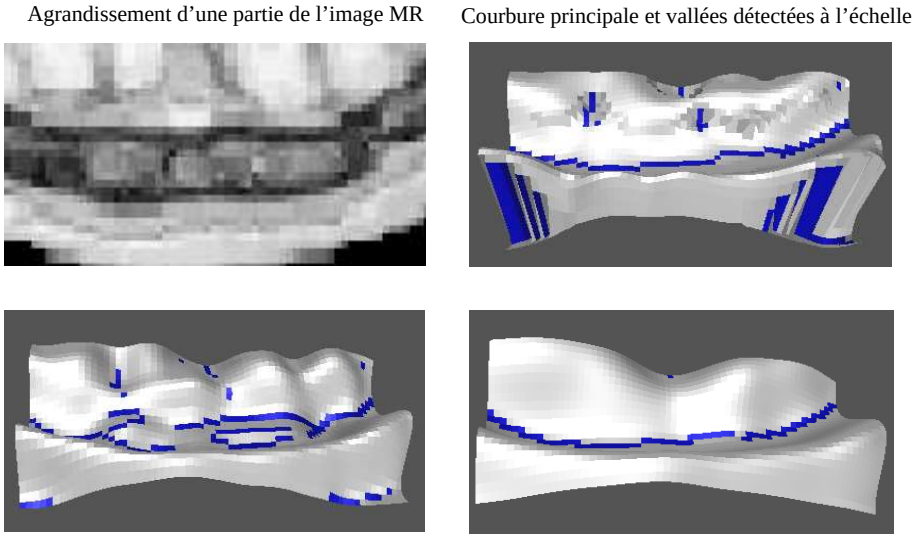


FIG. 2.21 – L'agrandissement d'une partie de l'image MR et la représentation 3D de la surface de courbure principale, Laplacien à l'échelle $\sigma = 4$, et Laplacien à l'échelle $\sigma = 8$.

L'image CT est plus simple à détecter que l'image MR parce que le centre de crâne y est plus homogène et plus nette par rapport au fond. Dans ce cas, le Laplacien donne la plus jolie réponse. La crête est parfaitement continue et fine (largeur d'un pixel). La courbure principale produit parfois des petites branches qui ne sont pas enlevées par le seuil de l'énergie de Laplacien.

Il faut noter qu'à part la vallée correspondant au centre de crâne dans l'image MR ou la crête dans l'image CT, il y a d'autres vallées/crêtes juste près du contour à l'extérieur. Il s'agit de fausses crêtes/vallées. En utilisant la technique de fenêtrage, ces fausses crêtes/vallées sont entièrement enlevées (voir figure 2.22).

Images d'empreintes digitales. Ces dernières années, les systèmes biométriques basés sur la reconnaissance d'empreintes digitales sont développés significativement. Les activités sur ce sujet concernent les académiques aussi que les industriels dont la plupart des approches se basent sur la squeletisation de l'image et ensuite la mise en correspondance des points de terminaison ou bifurcation des squelettes [FKVL99, Hol92, JHPB97, Vaj00]. Ainsi la détermination de squelettes dans l'image d'empreinte digitale est la première étape dans le système de vérification.

L'image d'empreinte digitale contient plusieurs traits "gris foncés" et "gris clairs" qui peuvent être représentés par les crêtes et vallées arrangées alternativement. Il est possible, au lieu de déterminer les squelettes ce qui nécessite une phase de binarisation, de détecter des crêtes et vallées et puis réaliser la mise en correspondance exactement comme avec les squelettes. Dans [BBD+02], Bishnu *et al.* ont classifié les pixels dans l'image d'empreinte digitale en 3 classes : crête, vallée, "slope line". L'extraction de crête utilisée dans ce travail est basée sur le test de changement de signe de Gradient. Cette approche

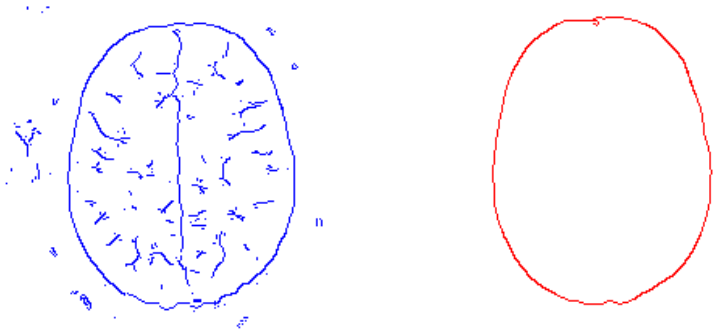


FIG. 2.22 – La détection de vallée et crête à partir de l'image MR et CT en utilisant le fenêtrage de Laplacien. Toutes les fausses crêtes/vallées sont éliminées. Le seuil utilisée pour l'image MR est 0.0, pour l'image CT est 19.0

fournit des crêtes épaisses qui nécessite une phase d'amincissement.

Nous testons notre algorithme de détection de crêtes par 3 mesures sur un ensemble d'images d'empreintes digitales fournies par **F**ingerprint **V**erification **C**ompetition[6]. La figure 2.23 montre le résultat de détection d'une image prise par un senseur optique à bas coût[7] à l'échelle $\sigma = 2$. La courbure de courbe de niveau donne plusieurs artefacts. Le Laplacien et la courbure principale détectent des crêtes continues. Une propriété remarquable de ces deux mesures est qu'elles détectent aussi bien des crêtes correspondant aux traits très floues avec une variation légère de l'intensité même difficiles à reconnaître par l'œil humain.

Images naturelles Nous réalisons d'autres tests de détection de crête sur les images naturelles comme celles de vélo, d'arbre, etc. Une même remarque est que le Laplacien et la courbure principale donnent un résultat similaire (voir 2.24). La courbure de la courbe de niveau donnent toujours des crêtes non significatives, discontinues.

Le test sur les images naturelles où les objets sont de formes et de tailles très variées montre l'expressivité de la crête pour la représentation des structures dans l'image. Le vélo, l'arbre sont des objets constitués de plusieurs tranches allongées. Il est préférable de les représenter symboliquement par des crêtes principales (vélo) ou réseau de crêtes (arbre) que par un certain nombre de points d'intérêt qui ne sont pas parfois les caractéristiques intrinsèques de l'objet. Une remarque importante est que la crête est présente dans tous les types d'images. Dans le cas d'une structure allongée, la crête correspondante est longue. Dans le cas d'une structure ronde, la crête est courte, considérée comme "pic" si le ratio des deux valeurs propres est inférieur à un seuil. Dans le cas où l'image est trop texturée, il n'y a pas de connexité entre des points de crêtes, les points détectés peuvent être considérés comme un type de "point d'intérêt".

2.6.3 Crêtes à l'échelle optimale et à multi-échelles

Nous venons de voir des exemples dans lesquels les crêtes sont détectées à une seule échelle pour représenter des structures d'une certaine taille dans l'image. Nous considérons maintenant le problème

[6] http ://bias.csr.unibo.it/fvc2004/

[7] Low-cost Optical Sensor

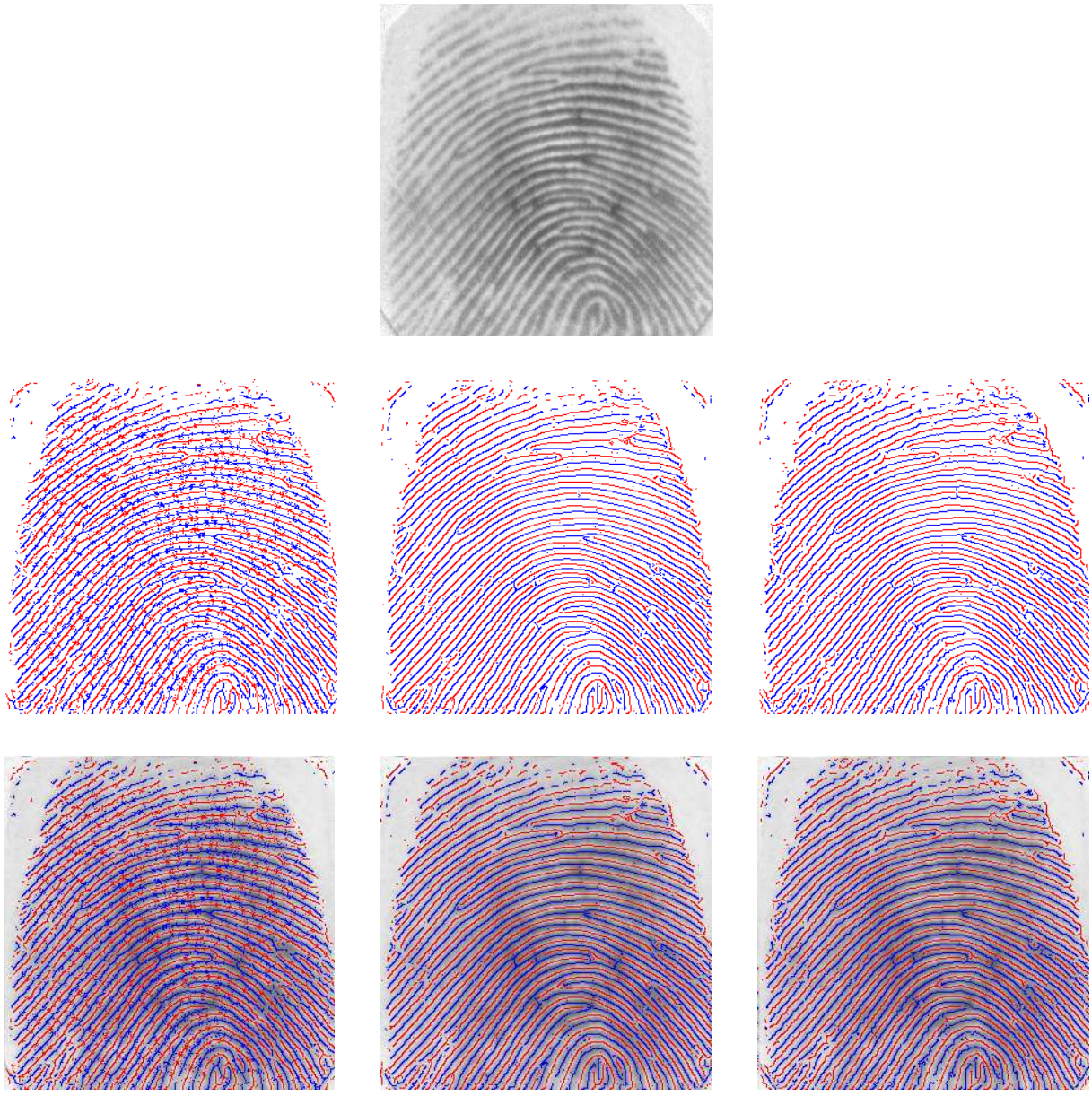


FIG. 2.23 – Détection de crêtes et vallées à l'échelle $\sigma = 2$ par 3 mesures : courbure de courbe de niveau, courbure principale, Laplacien.

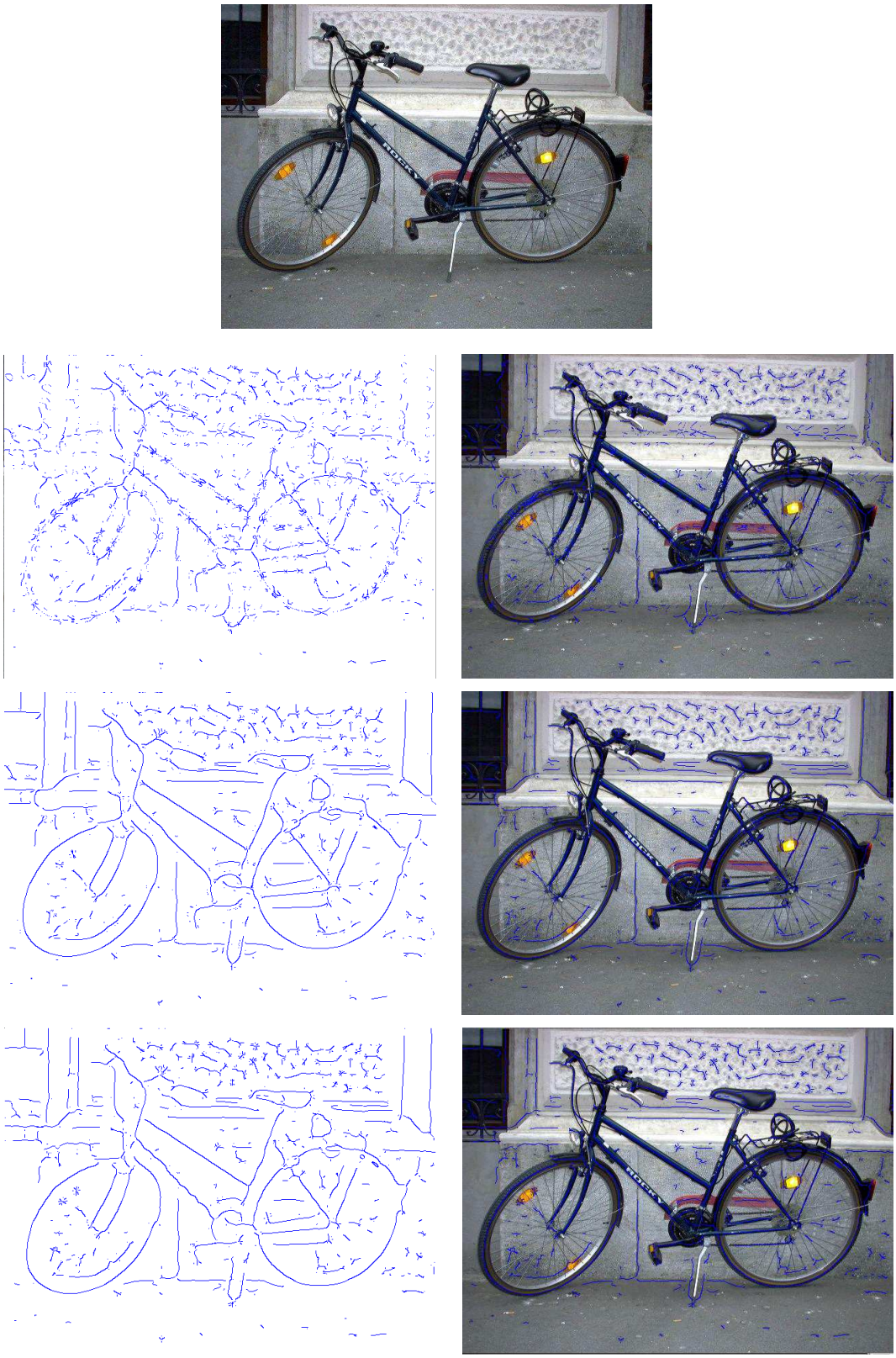


FIG. 2.24 – Crêtes détectées l'échelle $\sigma = 4\sqrt{2}$ par trois mesures : la courbure de la courbe de niveau κ, la courbure principale λ_1 et le Laplacien Lap.

de représenter un objet dans l'image. L'objet peut être constitué à partir de plusieurs éléments de taille différente.

Les sous-sections ci-dessous présentent deux approches pour décrire un objet dans l'image à base de crête : (1) à l'échelle optimale ; (2) à plusieurs échelles. Nous allons montrer quelques exemples où la première ou la deuxième approche va donner la meilleure représentation.

Détection des crêtes à l'échelle caractéristique.

Les images que nous étudions dans les sections précédentes contiennent des structures de largeur assez uniforme le long des structures (traits des caractères, cadre du vélo, branches de l'arbre, centre de crâne, etc). Il suffit donc dans ces cas de détecter des crêtes à une seule échelle. En réalité, le long une structure, la largeur de celle ci peut varier considérablement (frange de zèbre, coin d'une feuille). Ainsi, l'échelle caractéristique des points sur la crête correspondante varie aussi. La première façon de représenter de telles structures est de détecter chaque point de crête à son échelle caractéristique et ensuite enchaîner ces points pour former les lignes.

La détection des points de crêtes à l'échelle caractéristique est réalisée de la manière suivante :

- Calculer les dérivées L_x, L_y, L_{xx}, L_{yy}, L_{xy} et les mesures de crêtes tels que le Gaussien, la courbure, le Laplacien, etc à plusieurs niveaux.
- Calculer l'échelle caractéristique de chaque point dans l'image en vérifiant si le Laplacien admet un extremum local dans la dimension d'échelle.
- Pour chaque pixel, vérifier si les deux critères (2.12) sont satisfaits à l'échelle caractéristique de ce point.

Les figures 2.25 et 2.26 montrent des exemples dans lesquels les crêtes et vallées sont détectées à l'échelle caractéristique. Dans les figures, chaque point a une échelle caractéristique qui est représentée par une couleur. Nous constatons que le long des franges du zèbre, la largeur varie et cette variation est représentée par un changement de couleur sur l'image 2.25b. Le résultat de la détection des points à l'échelle caractéristique par le Laplacien fenêtré est représenté en bas de la figure.

Les crêtes et vallées détectées à l'échelle caractéristique sont remarquablement continues. En fait, il n'y a pas de difficulté dans la détection car la variation de l'échelle caractéristique le long de frange est continue. Il n'y a pas de changement brusque de l'échelle. Idem pour l'image de la feuille. La localisation de crêtes et vallées est correcte et très satisfaisante.

Détection de crêtes à plusieurs échelles

Un point peut avoir plusieurs échelles caractéristiques. Un exemple simple est la structure textuelle. Un point point se trouve à l'intersection entre l'axe médian du texte et celui d'un trait constituant un caractère a deux échelles caractéristiques, une est celle de caractère et une autre est celle de texte. L'approche de détection utilisant une seule échelle caractéristique n'est pas capable de représenter ces structures.

La figure 2.27 présente la détection de crêtes d'une image contenant le logo FOSTER'S. La détection est basée sur le Laplacien de Gaussian fenêtré calculé aux échelles $\sigma_1 = 4$ et $\sigma_2 = 16$. Plusieurs crêtes verticales représentent les squelettes des caractères à petite échelle tandis qu'à échelle plus grande, une seule ligne de crête correspond à l'axe médian de texte. Nous utilisons ces propriétés pour une application de détection de texte dans le chapitre 5.

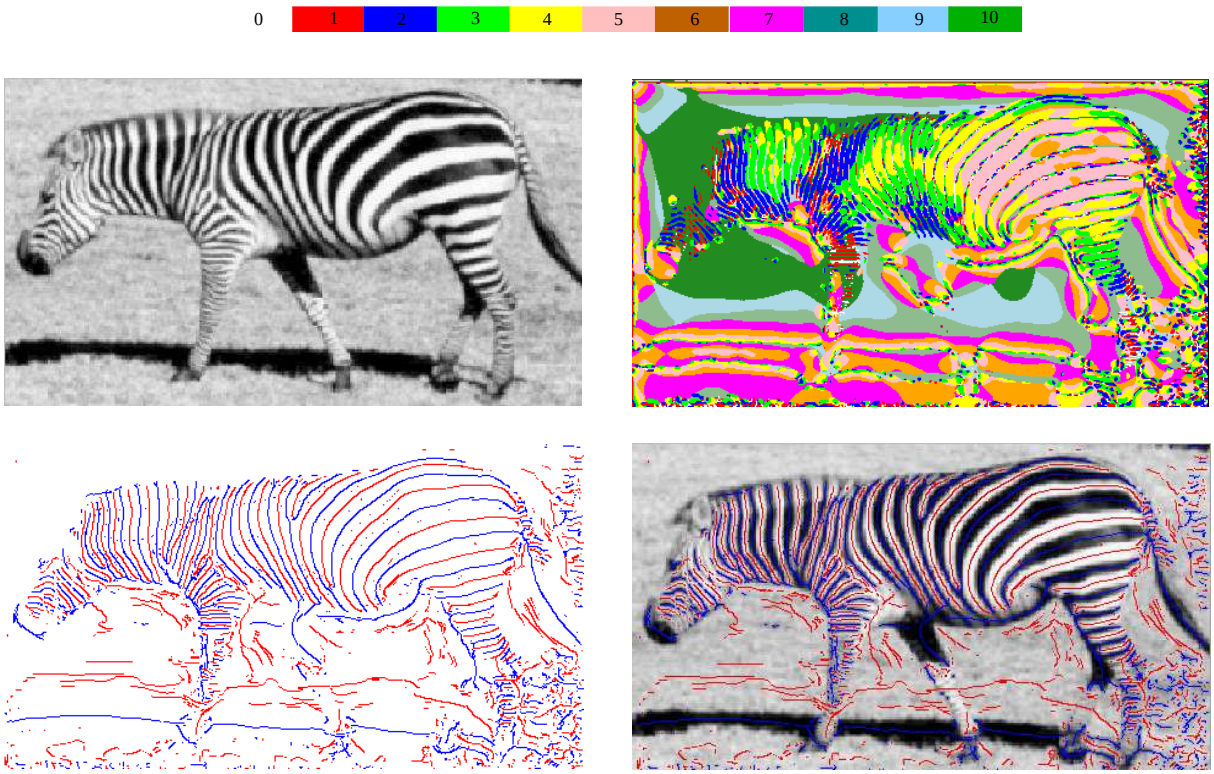


FIG. 2.25 – En haut : image d'un zèbre et les échelles caractéristiques de chaque pixel en pseudocouleur. En bas : détection de crêtes et vallées à l'échelle caractéristique avec le Laplacien de Gaussien fenêtré (le Laplacien est seuillé à 10.0).

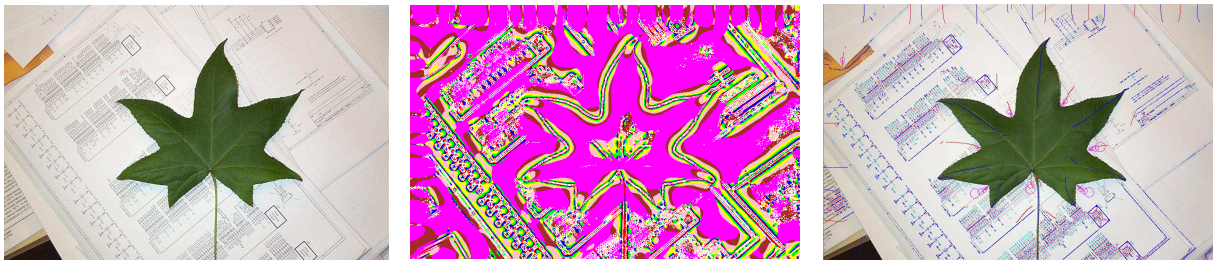


FIG. 2.26 – De gauche à droit : image d'une feuille, image en pseudocouleur avec l'échelle caractéristique de chaque pixel, superposition des crêtes, vallées, pics, trous détectés à leurs échelles caractéristiques avec le Laplacien du Gaussien.

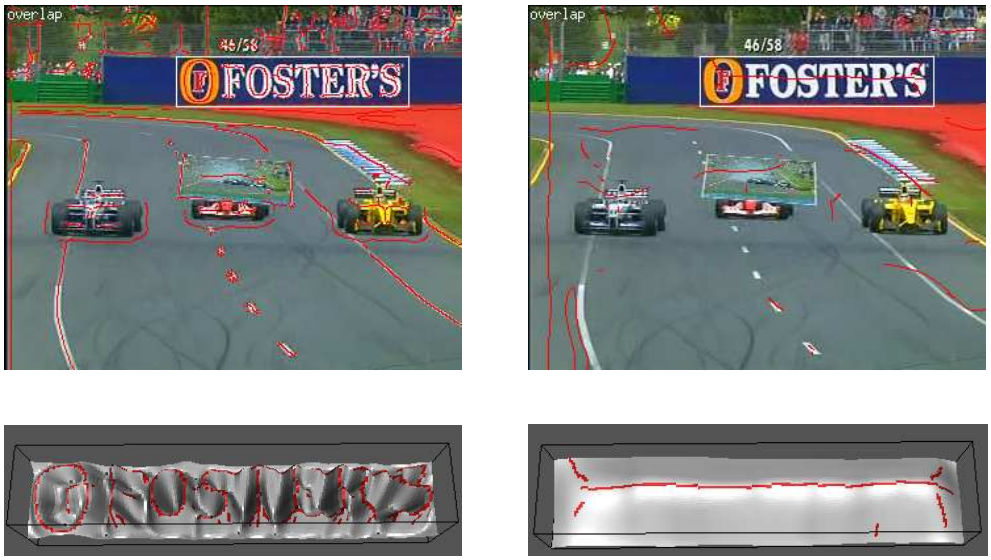


FIG. 2.27 – En haut : Image originale avec des crêtes en rouge. A gauche : à l'échelle $\sigma_1 = 4$, à droite à l'échelle $\sigma_2 = 16$. En bas : représentation 3D du Laplacien dans un rectangle couvrant le logo "FOSTER" à la même échelle que les crêtes.

La figure 2.28 montre les crêtes détectées à différentes échelles de l'image d'un arbre. Il est intéressante de remarquer que les petites branches disparaissent petite à petite jusqu'à moment où il ne reste que des grandes crêtes correspondant aux grandes branches et la tige.

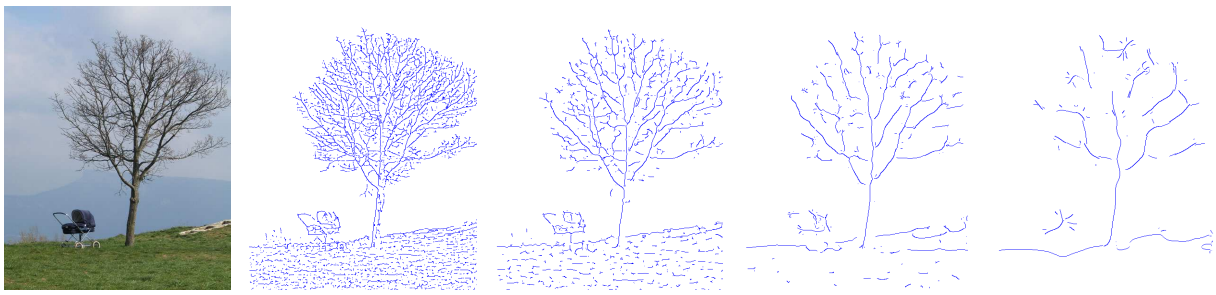


FIG. 2.28 – Image d'un arbre et les crêtes détectées à différentes échelles. De gauche à droite : $\sigma_1 = \sqrt{2}, \sigma_2 = 2, \sigma_3 = 2\sqrt{2}, \sigma_4 = 4$.

2.7 Conclusion

Ce chapitre a abordé le problème de détection de crêtes. Ci-dessous, nous résumons nos contributions principales et posons des problèmes ouverts à partir de l'étude de crête :

- D'abord, nous avons réalisé une étude de différentes définitions de crête. Ces définitions s'appliquent à des surfaces mathématiques continues. Pour les appliquer aux images, nous avons représenté l'image dans l'espace d'échelle, ajoutant un paramètre de lissage Gaussien de l'image originale. Les crêtes sont détectées sur les surfaces lissées.

  Pour détecter les points de crêtes, nous avons proposé une nouvelle définition basée sur le Laplacien. Cet opérateur connu pour la détection des points d'intérêt et des pics est aussi utile pour la détection de crêtes. Une évaluation expérimentale sur les images de textes dans différentes conditions a montré une bonne détection et localisation de crêtes détectées par le Laplacien.

  Parmi les trois mesures testées, la courbure de la courbe de niveau produit des crêtes moins continues. Le Laplacien et la courbure principale donnent un résultat comparable. Parfois le Laplacien détecte des crêtes plus continues et plus significatives s'il est calculé à l'échelle appropriée. Parfois, la discrétisation de l'espace de l'angle rend discontinue les crêtes détectés par le Laplacien. Un algorithme utilisant l'orientation exacte peut améliorer ce problème.

- Un inconvénient de notre calcul avec le Laplacien est l'apparition des fausses crêtes. Ce problème est commun à toutes les définitions basées sur la courbure. Nous avons proposé deux méthodes pour éliminer les fausses crêtes. Intuitivement, la méthode basée sur le passage par zéro est moins efficace que la méthode basé sur la technique de fenêtrage. En fait, la méthode basée sur le passage par zéro enlève à la fois des faux points et des vrais. La méthode basée sur le fenêtrage enlève des faux points de façon raisonnable. Pourtant le fenêtrage provoque des lobes supplémentaires dans le domaine fréquetiel de Laplacien du Gaussien. L'application de l'opérateur de Laplacien du Gaussien fenêtré sur l'image peut causer des éléments de haute fréquence qui peuvent être considérés comme des "nouveaux faux points de crête".

  En conclusion, nous trouvons que le problème de fausses crêtes est difficile à résoudre à l'origine parce que dans le cas où le fond est complexe, les structures de l'objet et celles du fond alternent et même l'être humain ne reconnaît pas les fausses des vraies. Nous pensons que ce problème peut être résolu par deux moyens :

  1. Si les crêtes ne sont pas significatives pour la représentation d'une classe d'objets (celles du fond par exemple), par un algorithme de clustering, ces crêtes seront classées dans une classe de minorités. De telles classes de caractéristiques ne sont pas utilisées pour la modélisation de l'objet.
  2. Nous constatons qu'en un vrai point de crête, le maximum de Laplacien existe à plusieurs échelles. En un faux point, le maximum n'existe qu'à une seule échelle. Autre interprétation, la position de fausse crête change dans deux échelles consécutives. Nous pouvons utiliser cette propriété pour détecter les fausses crêtes et des vraies.

- Une expérimentation sur différents types d'image montre l'expressivité de la crête pour la description de la forme d'objet. Les objets que nous avons montrés sont constitués des structures allongées, qui sont représentées de façon significative et efficace par les caractéristiques crêtes. Les points d'intérêt sont moins significatifs dans ce cas. Nous insistons sur le fait que une seule caractéristique de type crête ne permet pas de représenter tous les types d'objet. Par exemple dans

le cas où les objets sont texturés, les crêtes sont des points isolés aléatoirement arrangés. Cependant, pour un très grand nombre d'objets naturels, des structures allongées sont présentes, parfois déformés en pics. Dans ce sens, la caractéristique de type crête est une caractéristique générique pour la représentation de l'objet.

- Le problème de la représentation d'objets à base de crête pose la question de choisir entre deux approches : l'une se base sur l'échelle caractéristique et l'autre se base sur un nombre fixe d'échelles. Idéalement, un modéliseur doit faire sa décision automatiquement. Pourtant, ce choix est difficile sans assistance de l'être humain. Une idée serait de travailler à plusieurs échelles et d'enlever les réponses répétitives par utiliser une mesure de signification de crête en fonction de longueur, direction, couleur, etc.

Chapitre 3

Modélisation et classification de personnes

La détection automatique des personnes et la localisation des parties du corps sont les problèmes importants et provocateurs en vision par ordinateur. Ces problèmes se trouvent dans plusieurs applications telles que la navigation automatique des robots, la vidéo surveillance, l'interaction homme-machine, l'évaluation des performance des athlètes et l'animation.

La reconnaissance générique d'objet en général est un problème difficile à cause de la variation de luminosité, d'échelle et du point de vue. La reconnaissance d'une personne en particulier est encore plus laborieuse. Les difficultés proviennent du nombre de dégrés de libertés des parties du corps, de l'occlusion, de la variation de l'apparence à cause de vêtement et l'ambiguïté dans la projection d'une personne 3D au plan image. En outre, la région correspondante à une personne dans une image est souvent petite qui n'est pas suffisamment distinctive pour pourvoir appliquer la reconnaissance de visage ou de la main.

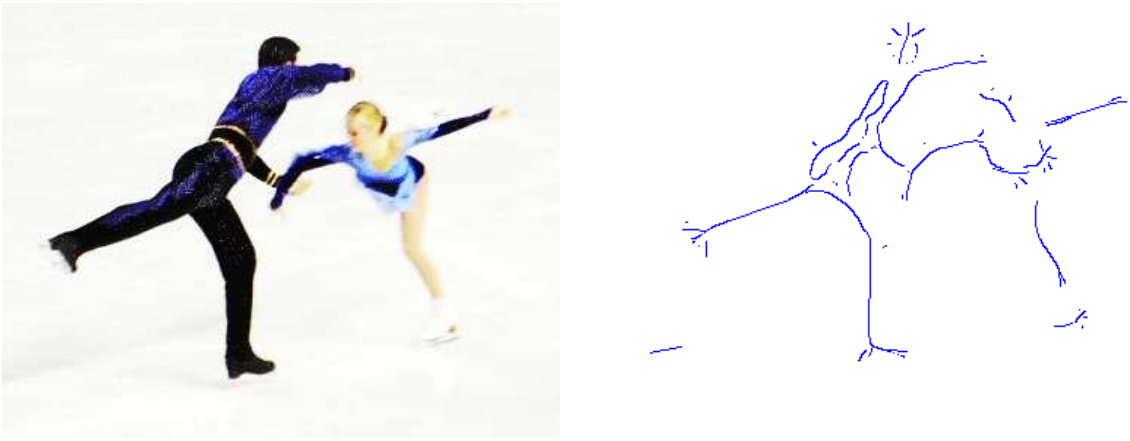
Les méthodes de détection de personnes en littérature utilisent souvent le contour[1] [BY92, Lai94, BH93, BH94, HHD98a, HHD98b, Har99, Zha01] ou les squelettes[2] [ZY96]. Ces méthodes ont beaucoup de difficulté en travaillant avec l'image des personnes dans un environnement réel. En fait, la détection de contour parfait et complet de l'objet d'intérêt reste un grand défis, surtout quand la personne se trouve dans une scène encombrée avec la variation de texture et la lumière (voir un exemple dans la figure 3.3).

Ce chapitre présente une application de crête à la modélisation et à la reconnaissance des personnes dans une séquence de vidéo. L'équipe PRIMA a développé une méthode fiable et stable pour suivre des objets mobiles dans une séquence vidéo [PC02b, PC02a, CHRC04]. A la suite du suivi, les régions mobiles doivent être reconnues pour les traitements ultérieurs tels que l'analyse de mouvement ou la reconnaissance de contexte.

L'objectif de notre travail est de déterminer s'il existe une ou plusieurs personnes dans une région donnée. La reconnaissance exacte n'est pas appropriée dans ce cas parce que la classification de personne et non-personne a besoin d'une représentation générique qui tolère une variation forte entre des personnes. En outre, nous voulons construire une représentation sémantique qui permet de mieux comprendre la configuration de la personne et aider à l'interprétation de son mouvement.

La méthode de représentation de personne que nous proposons se base sur l'observation qu'une personne peut être caractérisée par quelques crêtes principales correspondant aux structures telles que le

[1] silhouette
[2] stick figure



(a) (b)

FIG. 3.1 – (a) Image originale. (b) Vallées détectées à l'échelle $\sigma = 8$. Les vallées significatives représentent bien la configuration des personnes.

torse, les jambes, les bras. La détection de crêtes à l'échelle reliée à la taille des structures permet de représenter significativement la configuration de la personne (voir par exemple la figure 3.1). Chaque personne sera représentée par un vecteur de descripteurs qui préserve l'information géométrique structurelle du corps et la relation spatiale entre les structures.

L'organisation de ce chapitre est la suivante : La section 1 explique le contexte de travail, plus concrètement, l'entrée et le résultat à atteindre. La section 2 présente la classification d'objets mobiles en littérature. Nous proposons dans la section 3 une méthode de modélisation de personne basée sur les crêtes significatives. Cette méthode est validée sur les données réelles acquises dans la hall d'entrée de l'INRIA et comparée avec deux autres méthodes statistiques développées dans l'équipe PRIMA.

3.1 Contexte

Le projet PRIMA a développé une méthode de détection et de suivi des objets mobiles dans une séquence vidéo. Cette méthode est basée sur la soustraction de l'image courante avec l'image de fond (voir la figure 3.2) et la mise à jour de l'image de fond par une technique de seuillage. Elle marche en temps réel et fournit un résultat de suivi satisfaisant [CHRC04].

A chaque image de la séquence, le suivi fournit une liste de régions mobiles (régions d'intérêt[3]). On souhaite à déterminer à quelle classe appartient l'objet qu'elle contient. Dans une application de surveillance ou d'analyse de comportement des clients, les objets qui nous intéressent sont les "personnes". Le premier niveau de traitement est de classifier des objets en deux classes : personne et non-personne. Cette tâche s'inscrit dans le contexte du projet CAVIAR pour le test de surveillance dans le centre ville et d'analyse du comportement des clients.

[3]ROI-Region of Interest.

(a) (b) (c)

FIG. 3.2 – (a) Image courante. (b) Image de fond. (c) Image de soustraction. Les pixels jaunes sont les pixels de différence. La région contenant ces pixels est la région d'intérêt.

Supposons que la détection de région d'intérêt soit effectuée, notre système de classification des personnes prend comme entrée la région d'intérêt avec les caractéristiques telles que la position (x, y) du centre de gravité, la largeur et la hauteur (w, h) du rectangle qui l'encadre, l'orientation de l'axe principale (voir la figure 3.11). Le système vérifie si la région contient une personne ou pas. S'il y a des personnes il est préférable de dire combien.

3.2 Classification d'objets mobiles

La classification d'objets mobiles a été moins abordée que la reconnaissance générique d'objets. Ceci vient de trois raisons principales. D'abord, dans une application concrète, les catégories d'objets sont habituellement peu nombreuses : la surveillance étudie seulement deux classes d'objets : véhicule et personne. De plus, ces types d'objets sont prédéfinis et donc nécessitent des techniques de modélisation plus particulières. Enfin, la classification doit être simple pour que le système marche en temps réel. Les approches de classification sophistiquées comme présentées dans le chapitre 2 n'apparaissent pas appropriées.

La plupart des méthodes de classification de la littérature utilisent les informations statiques de la forme telles que la compacité (le ratio du périmètre et de la superficie), le rapport entre les longueurs de deux axes principaux, etc. Ces mesures simples, faciles à calculer et montrées efficaces dans certains cas où la taille des objets ne change pas fortement dans une séquence et qu'il n'y a pas de distorsion de la forme de l'objet.

Dans [LFP98], Lipton *et al.* ont intéressée à trois classes d'objets : la personne, la voiture et l'animal. La détection de ces objets s'effectue en soustrayant l'image courante de l'image de fond pour déterminer les régions de mouvement. La classification est réalisée en utilisant une mesure de disparité de la région de chaque objet. Le choix de cette mesure est part de l'observation que la personne par sa forme complexe et sa petite taille, a la disparité plus grande que celle de la voiture ou de l'animal [LFP98].

La méthode de Lipton *et al.* nécessite la détection de contour externe de l'objet. Ceci n'est pas toujours évident (voir figure 3.3). Premièrement, l'objet dans la scène se déplace, son contour devient flou. Deuxièmement, si une partie de l'objet a la même couleur que celle du fond, la région détectée contient des trous. Troisièmement, l'ombre peut fausser la région de détection de l'objet.

(a) (b) (c)

FIG. 3.3 – (a) Image courante contenant l'objet mobile. (b) Détection de région de mouvement par la soustraction. (c) Les contours de la région de mouvement vus à la loupe.

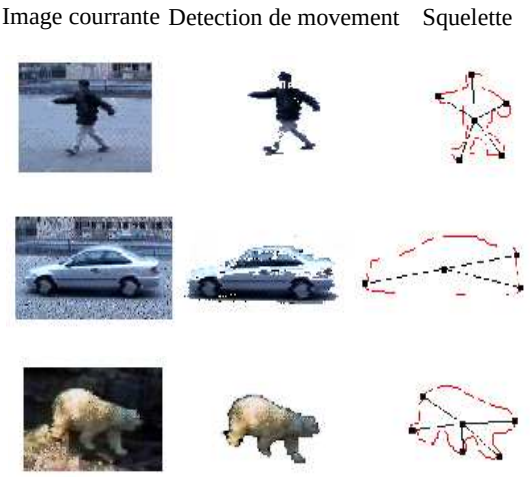


FIG. 3.4 – Modélisation des objets mobiles par les réseaux de squelettes [FL98].

Supposons que le contour externe soit déterminé, la disparité apparaît trop simple pour distinguer des objets. Lipton *et al.* ont amélioré leur représentation par un réseau de squelettes construit à partir de son contour externe (voir la troisième colonne de la figure 3.4). Cette représentation permet de reconnaître plus fiablement les objets et d'interpréter leurs mouvements [CLK+00].

3.3 Modélisation de personne à base de crêtes principales

Dans cette thèse, nous nous intéressons à la classification de deux classes d'objets : personne et non-personne. Les objets tels que la voiture ou l'animal, etc. sont tous classés en classe non-personne. Ainsi, nous voulons détecter les caractéristiques plus distinctives de personnes pour les distinguer avec des autres types d'objets "non-personnes". La modélisation de la personne sera étudiée plus profondément dans cette section.

Les travaux de la littérature concernant la modélisation de personne dans une séquence vidéo utilisent les caractéristiques telles que le contour [BH94, ZNL01], le squelette [ZLK04] ou l'étoile de squelettes[LFP98, FL98], etc. Ces caractéristiques sont manuellement ou automatiquement détectées. Avec tous les inconvénients que l'on trouve à la détection automatique de contour ou squelette, nous proposons d'utiliser des crêtes.

La détection de crête est en effet plus robuste que la détection de contour ou de squelette. Ensuite, une crête représente une structure de l'objet. Une personne se compose de structures telles que la tête, le torse, les bras et les jambes. Ces structures sont longues, généralement assez uniformes en couleur (sauf quand la personne porte des vêtements à fleurs). Elles peuvent donc être représentées par des crêtes à l'échelle reliée à la taille de structure.

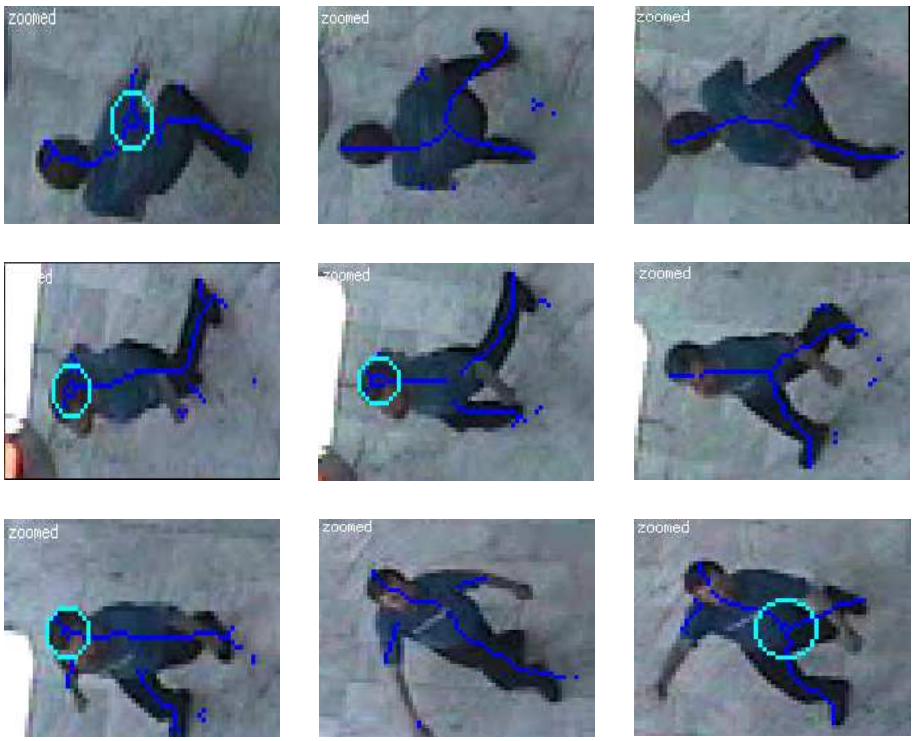


FIG. 3.5 – Différentes configurations de la personne représentées par crêtes (lignes bleus) et pics (cercles cyans) à l'échelle $\sigma = 4\sqrt{2}$.

La figure 3.5 montre un exemple dans lequel une personne apparaît dans plusieurs images d'une séquence vidéo. Les crêtes et pics détectés à l'échelle $\sigma = 4\sqrt{2}$ superposées sur l'image originale montrent leurs correspondances très satisfaisantes avec les membres de la personne dans l'image. L'ensemble de crêtes significatives représentent effectivement la configuration de la personne. Nous proposons donc une méthode de modélisation de personne à base de crêtes.

Notre système de modélisation de personne utilise la région de mouvement détectée par le suivi. Il extrait des crêtes à quelques échelles appropriées dans la région d'intérêt. L'orientation de la personne dans la région est ensuite déterminée qui permet de séparer la région en deux parties : la partie corres-

pondant au torse et la partie correspondant aux jambes. Les crêtes les plus significatives dans chaque partie sont utilisées pour modéliser la personne.

3.3.1 Détection de crête dans la région de mouvement

La première étape est d'extraire des crêtes dans la région d'intérêt. La région d'intérêt est définie par un rectangle (x, y, w, h, θ) où (x, y) est la position du centre de gravité, w, h sont la largeur et la hauteur du rectangle et θ est l'orientation de l'axe principal de la région.

L'algorithme de détection de crête à l'échelle σ est appliqué sur la région d'intérêt. Une question qui se pose est quelle est l'échelle à laquelle on détecte des crêtes ? Nous supposons que la région d'intérêt couvre parfaitement la personne, l'échelle pour la détection de crête représentant le torse est alors égale à la moitié de la largeur w et celle représentant les jambes est égale à un quart de w.

L'expérimentation de détection de crête sur la région correspondant à une personne nous montre qu'avec l'utilisation de Laplacien, quelques crêtes représentant une même structure se répètent à plusieurs échelles. Ceci arrive également dans le cas d'une personne : les crêtes détectées à l'échelle du torse dans la partie jambe représentent bien les jambes. Nous proposons de détecter des crêtes à une échelle unique qui est l'échelle du torse ($w/2$) et également l'échelle caractéristique de la personne.

3.3.2 Détermination de crêtes principales correspondant au torse et jambes

Quand la personne est debout, la crête correspondant au torse passe près du centre de gravité de la région et parallèle à l'axe principal le plus long de la région. Malheureusement, ceci n'est pas toujours le cas. Il est probable que le torse n'est pas alligné à l'axe principal de la région et la crête correspondante peut ne pas passer le centre de gravité. La détermination de crête correspondant au torse et aux jambes n'est pas donc évidente en général.

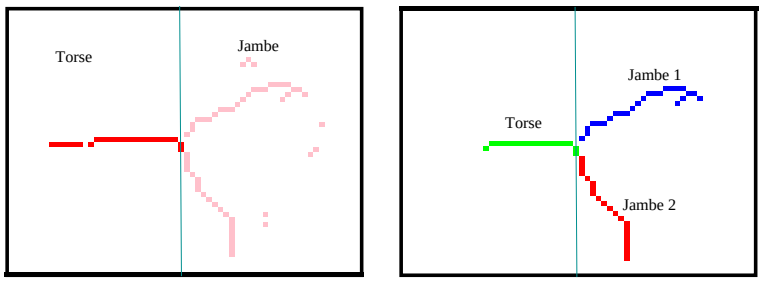


FIG. 3.6 – Détermination de crêtes principales correspondant aux torse et jambes. (a) Séparation des deux parties jambes et torse. (b) Identification de crête correspondant aux torse et jambes.

Nous supposons connaître l'orientation ϕ de la personne. Ainsi, nous séparons la région d'intérêt en deux parties torse et jambe par un axe passant le centre de gravité et perpendiculaire à l'axe aligné à l'orientation de la personne (la ligne cyan dans la figure 3.6). Les crêtes significatives en longueur et énergie sont considérées (ie. les crêtes ayant la valeur de Laplacien et de longueur supérieure à un seuil).

La crête la plus longue dans la partie torse représente le torse. Deux crêtes les plus longues dans la partie jambe représentent les deux jambes.

Il se peut qu'il n'y ait pas de crête torse ou il y a une seule ou zéro crête jambe. Ceci arrive quand la personne porte des vêtements à fleur ou à rayures où les structures longues ne se présentent pas ou dans le cas où la personne est cachée partiellement, la crête longue est coupée en deux crêtes plus courtes. Ceci n'est pas grave parce que ça permet de différentes configurations et rend le modèle robuste à l'occultation. En utilisant des crêtes pour la modélisation, une personne se trouve dans une des configurations montrées dans la figure 3.8.

3.3.3 Construction des descripteurs

Nous représentons une configuration de personne par un vecteur de 10 composantes déterminées à partir de trois crêtes principales.

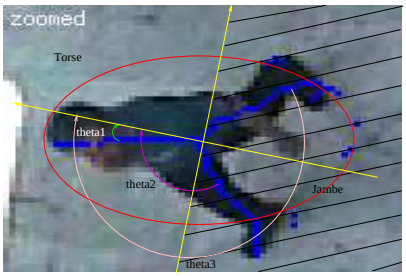


FIG. 3.7 – Une personne est modélisée par un vecteur de 10 composantes.

$$descripteur = (n, \theta_1, len_1, dis_1, \theta_2, len_2, dis_2, \theta_3, len_3, dis_3)$$

Le premier élément est le nombre de crêtes "significatives" détectées dans la partie torse et la partie jambe de la région d'intérêt. Ce nombre peut être 0, 1, 2, 3 :

- $n = 0$, la région ne contient aucune crête significative. Cela signifie l'absence de la personne dans la région.
- $n = 1$, une seule crête est détectée dans la région d'intérêt. Elle peut correspondre au torse ou à une jambe.
- $n = 2$, deux crêtes sont détectées. Elles se peuvent (1 crête torse + 1 crête jambe) ou 2 crêtes jambes.
- $n = 3$, il y a une crête torse et deux crêtes jambes. Cette configuration est la plus complète.

Nous constatons que le cas où $n = 2$ ou 3 ne correspond pas à une configuration unique. Par conséquent, nous proposons d'assigner à chaque crête un poids selon son importance (ie. 1 pour crête jambe et 3 pour crête torse). Alors, n est la somme pondérée des crêtes détectées. Il devient $\{0, 1, 2, 3, 4, 5\}$. Chaque index correspond à une configuration unique.

Les 9 composantes restantes dans le vecteur de descripteurs sont 3 triples l'angle entre crête et l'axe principale la plus longue, la longueur de la crête et la distance entre le centre de gravité de la crête et

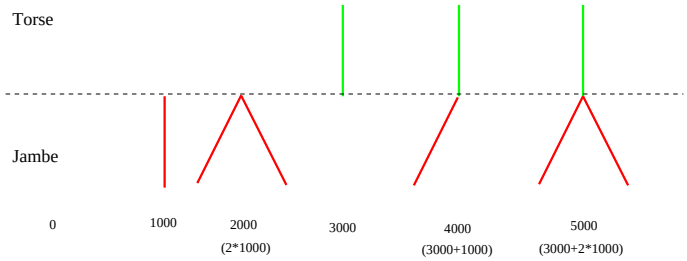


FIG. 3.8 – 5 configurations

celui de la région d'intérêt. La longueur des crêtes est normalisée à l'échelle pour rendre le descripteur invariant au changement d'échelle. Tous les composantes sont ensuite normalisées par leur valeurs maximales qui sont expérimentalement choisies : $\theta_{max} = 2\pi$, $len_{max} = 35$, $dis_{max} = 17$.

Parmi 10 composantes du vecteur de descripteurs, la première est la plus importante parce qu'elle informe de la configuration de la personne. Ainsi nous mettons un poids important sur cette composante (1000 dans notre expérimentation). Après avoir classifié un modèle de personne en une de 5 classes, nous déterminons plus précisément la relation entre des crêtes par 9 composantes restantes.

3.4 Apprentissage des modèles de personne

Les crêtes correspondant aux torse et jambes d'une personne représentent la forme 2D de la personne. La reconnaissance d'une personne dans une région donnée consiste à vérifier si le modèle construit dans cette région est similaire à un des modèles de personnes appris.

La façon la plus simple pour apprendre les modèles de personne est de stocker tous leurs vecteurs de descriptions. La reconnaissance revient à chercher le vecteur le plus proche du vecteur donné. Le fait de stocker tous les modèles appris est simple mais plusieurs inconvénients sont présents. D'abord, le nombre important de modèles implique un temps de recherche considérable. Ensuite, les modèles stockés sont les configurations précises des personnes. Ils sont moins tolérants à la variation de l'apparence de personne.

Pour accélérer la mise en correspondance et tolérer la variation entre les entités dans une classe, nous utilisons l'algorithme de classification K-Means pour classifier des modèles appris. K-Means requiert la spécification du nombre de classes. Comme nous ne connaissons pas le nombre classes suffisantes pour représenter toutes les configurations de personne, nous testons avec plusieurs nombres $n_s = \{8, 12, 17, 22, 31, 34\}$.

L'expérimentation montre qu'avec $n_s = 34$, toutes les configurations significatives de personne sont présentes (voir la figure 3.9). Nous choisissons alors $n_s = 34$ pour le test de classification. Dans cette figure, les lignes vertes représentent les crêtes torses, les lignes rouges représentent des crêtes jambes. Cette figure est juste pour l'illustration. En réalité, les crêtes ne coïncident pas avec le centre de la région.

Une remarque est que le cluster 2 ne contient aucune crête. La raison est que quelques régions d'intérêt fournies à l'apprentissage sont trop petites. La sélection de crête significative demande que la crête soit plus longue qu'un seuil. Cette contrainte élimine toutes les crêtes courtes dans les régions. Ainsi, un modèle appris ne contient aucune crête significative. De telles régions sont difficiles à reconnaître même

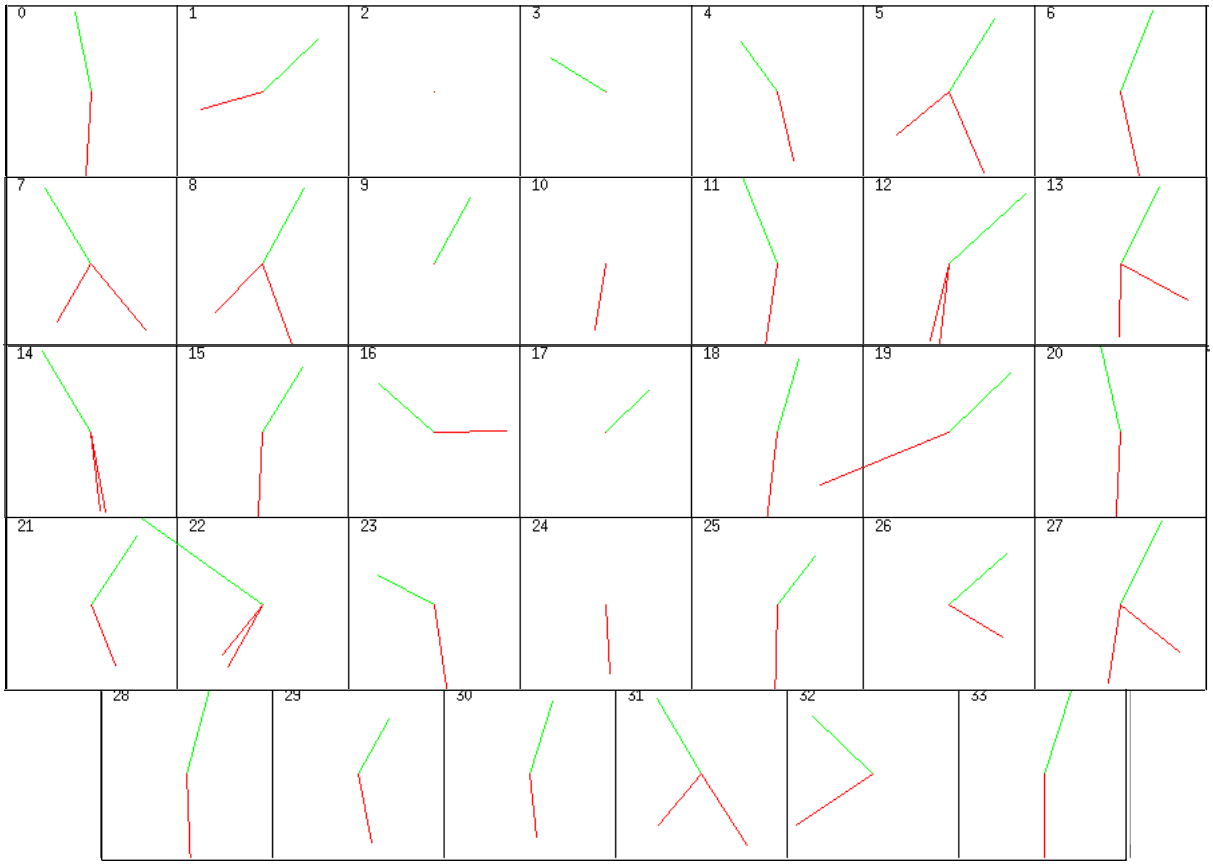


FIG. 3.9 – 34 clusters représentant 34 configurations significatives de la personne.

avec l'œil humain. Par conséquent, nous ne considérons pas ce cluster dans la suite.

3.5 Classification : personne et non-personne

Lors du traitement en ligne, le tracking nous fournit une région de mouvement à reconnaître. Nous appliquons les mêmes étapes de construction du "modèle de personne" : détection de crêtes à l'échelle appropriée, détermination de crête torse et crêtes jambes, construction du vecteur de descripteurs.

Dans la phase de test, nous ne connaissons pas l'orientation de la personne. Ainsi, pour chaque région, deux vecteurs de descripteurs sont construits correspondant à deux orientations torse-jambe et jambe-torse. Ces deux modèles sont comparés avec 33 modèles appris. La mesure de dissimilarité entre un nouveau modèle avec un modèle appris est la distance euclidienne entre deux vecteurs de descripteurs (cf. 3.1). La plus petite dissimilarité parmi 66 mesures (2 x 33) est choisie comme la meilleure

correspondance (figure 3.10).

$$dissimilarite_i = min\{\sum_{k=1}^{9} \sqrt{(descripteur_{jk} - descripteur_{ik})^2}, j \in \{1,2\}, i \in [1,33]\} \quad (3.1)$$

où i indique le $i$ème modèle appris, k indique le $k$ième composante du vecteur de descripteurs, j indique le modèle construit selon l'orientation torse-jambe ou jambe-torse.

FIG. 3.10 – (a) Une image de scène et la personne détectée. (b) Les crêtes significatives détectées dans l'imagette. (c) le modèle construit. (d) Le modèle le plus ressemblant dont l'index est $i = 31$. La dissimilirité est $d = 0.134369$.

La mise en correspondance donne un résultat composé de deux valeurs (i, d) signifiant l'index du modèle le plus ressemblant et la dissimilarité entre deux modèles. Plus la dissimilarité est petite, plus l'objet dans la région considérée est similaire à une des configurations de personne et plus fiable la reconnaissance de classe de personne. La classification détermine à la fois la présence d'une personne dans la région et son orientation.

3.6 Validation

Pour valider la performance de notre méthode, nous utilisons 24 séquences vidéos prises dans le hall d'entrée de l'INRIA fournies par CAVIAR. Les 12 séquences sont utilisées pour l'apprentissage de modèles de personne et les 12 restantes pour le test. Dans la phase d'apprentissage et la phase de test, nous utilisons les régions fournies par le groundtruth pour évaluer plus précisément. Le groundtruth ne fournit que les exemples de personne. Les exemples de non-personnes sont crées aléatoirement sur deux séquences de scène vide.

A chaque image de la séquence, le groundtruth fournit une liste d'imagettes qui sont des rectangles caractérisés par (x, y, w, h, θ) (voir la figure 3.11). Dans le groundtruth, il y a non seulement l'information d'une personne mais aussi de groupe de personnes. Dans ce cas, chaque personne dans le groupe est également décrite par ces données. L'index du groupe est fourni pour indiquer l'appartenance de

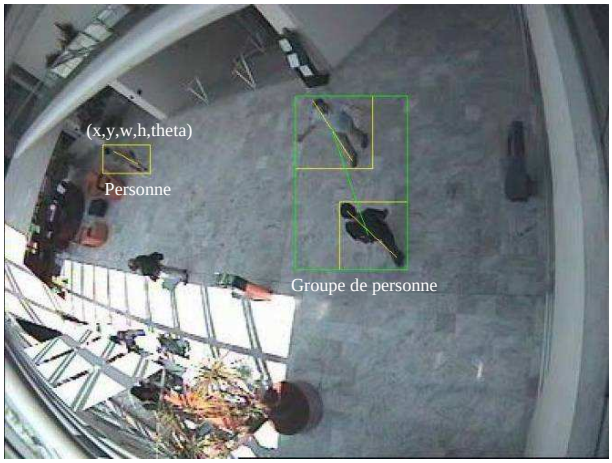


FIG. 3.11 – Un exemple de groundtruth

la personne dans quel groupe. Nous utilisons l'information de groupes des personnes pour le test de l'estimation du nombre de personnes dans un groupe, présentée dans la suite.

Les imagettes représentant les objets non-personnes sont caractérisées de la même manière. Elles se trouvent aux endroits dans la scène tels que le bar, le banc, le point d'information, la plante, le sol, etc. La taille de ces imagettes non-personne varient fortement. La figure 3.12 montre l'exemple d'objets non-personnes détectés sur une image de fond.

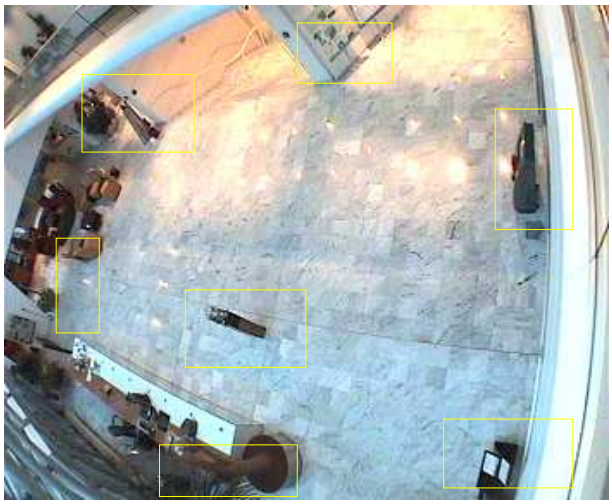


FIG. 3.12 – Un exemple de la création des imagettes non-personne.

L'apprentissage de modèles de personne à partir de 12 séquences fournit une base de 33 modèles (voir la figure 3.9). Pour le test, nous appliquons l'étape de classification personne et non-personne comme présentée dans la section 3.5. Un seuil est expérimentalement déterminé pour décider si l'objet dans la

région donnée est une personne. Plus concrètement, si $d < seuil_{person}$, l'objet est personne si non l'objet est non-personne.

La performance de la méthode de classification est évaluée par le rappel et la précision définis ci-dessous :

$$Rappel = \frac{\sum_{i=1}^{N_p} t(i)}{N_p}$$

où N_p est le nombre d'imagettes fournies par le groundtruth à partir de 12 séquences de test. Chaque imagette ne contient qu'une seule personne. Nous ne classifions pas les groupes de personnes.

$$t(i) = \begin{cases} 0 & \text{si } d_i > seuil_{person} \\ 1 & \text{si } d_i \leq seuil_{person} \text{ et } imagette_i \text{ contient une personne (confirmé par le groundtruth)} \end{cases}$$

$$Precision = \frac{\sum_{i=1}^{N_p} t(i)}{\sum_{j=1}^{N_p+N_{0p}} r(j)}$$

où N_{0p} est le nombre d'imagettes non-personnes crées aléatoirement.

$$r(j) = \begin{cases} 0 & \text{si } d_j > seuil_{person} \\ 1 & \text{si } d_j \leq seuil_{person} \end{cases}$$

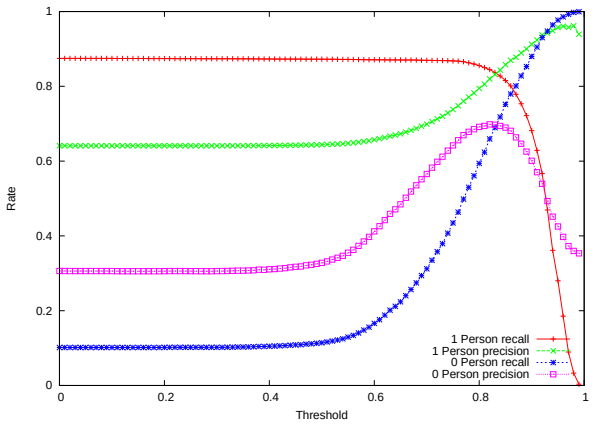


FIG. 3.13 – Résultat de classification personne et non-personne. L'axe horizontal représente la variation de $seuil_{person}$, l'axe verticale représente le taux de reconnaissance.

La figure 3.13 montre le résultat de reconnaissance des deux classes : personne et non-personne sur les données décrites. La mesure de dissimilarité est normalisée à 1 pour rendre le résultat comparable avec d'autres méthodes.

Nous constatons que la méthode proposée reconnaît bien les personnes. Le rappel est fiable (88%) et stable face à la variation de $seuil_{person}$ dans l'intervalle $[0, 0.7]$. Pourtant, la reconnaissance des non-personnes dans cet intervalle n'est pas satisfaisante (rappel est de l'odre 10%). La meilleure classification

de deux classes obtenue quand $seuil_{person} = 0.88$. En cette valeur, le rappel de classification de personne et non-personne est de 80%.

Les raisons pour lesquelles la méthode proposée n'est pas efficace à distinguer la classe non-personne sont multiples. D'abord, la méthode accepte des modèles ayant une seule crête. Ainsi toutes les régions contenant une seule crête peuvent être considérées comme régions de personnes. Or, cette situation arrive souvent. Ensuite, quelques imagettes crées aléatoirement pour représenter les non-personnes ont des caractéristiques de personne : la colonne d'information, la barre près du mur, etc. Enfin, le vecteur de description est discriminant pour les configurations de personnes. Pourtant, il ne l'est pas suffisamment pour distinguer une personne d'un objet non-personne.

3.7 Comparaison avec deux méthodes statistiques

Dans le cadre du projet CAVIAR, l'équipe PRIMA a développé deux méthodes de classifications de personne et non-personne [NTG+05]. Il s'agit de travaux de A. Nègre et N. Gourier. Dans la suite, nous présentons brièvement ces deux approches. Le but est de comparer la performance qu'elles obtiennent sur le même test réalisé avec la méthode proposée ci-dessus.

3.7.1 Reconnaissance basée sur l'histogramme du Gradient

L'idée principale de cette approche est de représenter une personne par une crête principale détectée dans l'espace d'échelle et de décrire cette crête par un histogramme de la magnitude et de l'orientation du Gradient. Cette approche est similaire à des approches basées sur l'histogramme de champs réceptifs[SC96c, SC00] ou le SIFT descripteur [Low04]. La construction du modèle de la personne se compose de 2 étapes :

Etape 1 : Détection de crêtes multi-échelle

La première étape consiste à détecter des points de crête dans l'espace d'échelle. Ensuite, les lignes de crêtes sont construites en analysant les composantes connexes dans l'espace (x, y, σ). Deux points de crête dans l'espace (x, y, σ) sont assignés à une même ligne de crête s'ils sont de même type (crête ou vallée) et la différence d'angles de deux points est inférieure à un seuil.

Chaque ligne de crête est en fait une composante connexe dans l'espace d'échelle. Elle est paramétrée par le centre de gravité μ pondéré par la valeur absolue de Laplacien normalisée et C_{ij} la matrice de covariance et σ_m l'échelle caractéristique moyenne. La matrice de covariance permet d'estimer la taille et l'orientation de la ligne de crête.

Etape 2 : Description statistique de la crête

Parmi des lignes de crêtes détectées dans l'étape précédente, la crête la plus significative est choisie. La signification de la crête est mesurée par l'énergie moyenne de Laplacien de tous les points appartenant à la ligne de crête.

En chaque point de la ligne de crête la plus significative, la magnitude et l'orientation du Gradient sont calculées. La magnitude est normalisée par la Gaussienne anisotropique $G(\sigma_1, \sigma_2)$ où $\sigma_1 = 2\sqrt{\lambda_1}$ avec λ_1 est la plus grande valeur propre de la matrice de covariance C_{ij} et $\sigma_2 = \sigma_m$ est l'échelle

caractéristique moyenne de la ligne de crête. Cette normalisation a pour but de se concentrer sur le point central de la crête. L'orientation du Gradient est calculée relativement par rapport à l'orientation θ de la région considérée.

Chaque région est modélisée par un histogramme de magnitude et d'orientation du Gradient. Pour des raisons de tolérance à la variation des objets inter-classes et de réduction du temps de recherche, les modèles sont également classifiées par l'algorithme K-Means.

Classification : personne et non-personne

La classification est réalisée en construisant l'histogramme de magnitude et d'orientation du Gradient de la région considérée. Cet histogramme est ensuite comparé avec ceux dans la base de histogrammes. La comparaison des histogrammes est effectuée en utilisant la distance χ^2-divergence.

Reconnaissance

La reconnaissance par l'histogramme du Gradient est évaluée de même façon avec la même base que la méthode précédente. La figure 3.14 montre la précision et le rappel de la méthode variant selon la valeur α qui est définie comme la probabilité d'occurrence de non-personne en général.

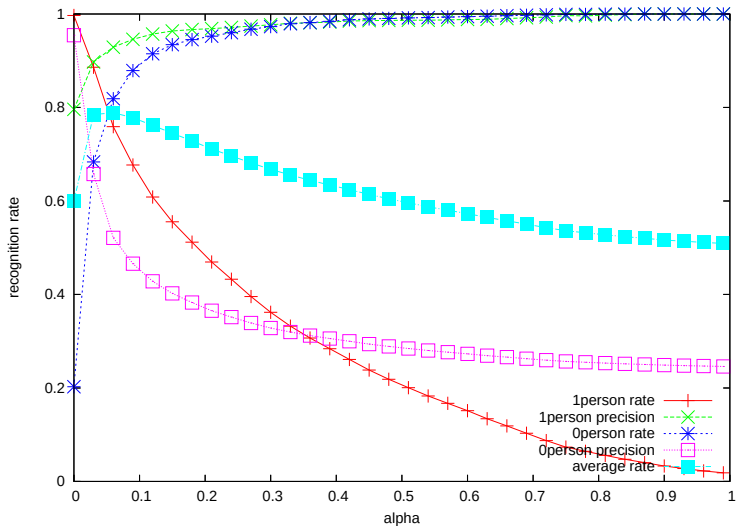


FIG. 3.14 – Résultat de classification en utilisant l'histogramme de magnitude et d'orientation du Gradient.

Nous constatons que le rappel se réduit significativement quand α augmente. La valeur α à laquelle le rappel de classification de personne et non-personne sont optimaux est égale à 0.09. Dans ce cas là, le rappel est d'environ 82%. Par rapport au rappel obtenu par la méthode 1, cette valeur est légèrement meilleure.

La méthode commet la même erreur de classification de non-personnes dans les cas où les imagettes représentant des non-personnes ont des caractéristiques de la personne par exemple la colonne d'infor-

mation, la barre près du mur, etc. A part ceci, cette méthode est sensible au changement de luminosité locale et de voisinage de l'objet.

3.7.2 Reconnaissance basée sur les mémoires auto-associatives

Si les deux méthodes de classification ci-dessus utilisent les informations structurelles, locales de l'objet pour la modélisation, la troisième approche utilise l'information globale, c-a-d l'imagette entière. L'idée est de représenter chaque classe d'objet par une matrice qui représente la connexion entre leurs entités. Cette matrice est corrigée petit à petit pour améliorer la possibilité de se distinguer avec d'autres classes.

Mémoires linéaires associatives

Les mémoires linéaires associatives sont un cas particulier de réseaux de neurones à une couche. Dans cette approche, chaque classe d'objet k est décrite par une matrice de connexion W_k entre ses entrées qui sont les imagettes dans ce cas. La matrice permet d'estimer l'imagette de sortie y étant donnée l'imagette d'entrée x : $y_k = W_k \cdot x$. La similarité entre l'image source x et classe k est $cos(x, y) = x \cdot y^T$.

Initialement, $W_k^0 = X_k \cdot X_k^T$ où $X = (x_1, x_2, ..., x_n)$ est une matrice nxm où n est le nombre de pixels dans chaque imagette x_i , $i \in [1, m]$, m est le nombre d'imagettes d'apprentissage. Dans le contexte de classifier les classes personnes et non-personnes, k ne vaut que deux valeurs représentant personne et non-personne. La matrice de connexion est corrigée à chaque itération par la règle de Widrow-Hoff :

$$W_k^{t+1} = W_k^t + \eta \cdot (x - W_k^t \cdot x) \cdot x^T$$

où η est le facteur d'apprentissage, t est le paramètre de temps. Dans l'expérimentation, $\eta = 0.06$ et le nombre d'itérations est environ 50.

Résultat de reconnaissance

Le taux de classification de la méthode est représenté par la figure 3.15. Sans apprentissage de la classe non-personne, cette méthode est inefficace à reconnaître les non-personnes. Un compromis de classification de deux classes obtenue est environ 25%. Ce rappel n'est absolument pas satisfaisant.

Le taux de reconnaissance est significativement amélioré avec l'apprentissage des objets non-personnes : 99% de reconnaissance correcte de classe personne et 68% de reconnaissance correcte de classe non-personne. Malgré l'utilisation de l'imagette entière, cette méthode de classification n'est pas capable de distinguer une colonne, une barre dans la scène avec une personne.

3.7.3 Conclusion sur trois méthodes de classification

Nous venons de présenter 3 méthodes de reconnaissance de personne : la première méthode basée sur la relation géométrique structurelle de quelques crêtes significatives (méthode 1), la deuxième méthode basée sur l'histogramme du Gradient de la crête la plus significative (méthode 2) et la troisième méthode basée sur les mémoires linéraires auto-associatives (méthode 3). Les deux premières méthodes sont locales tandis que la troisième est globale. La première méthode est structurelle tandis que les deux

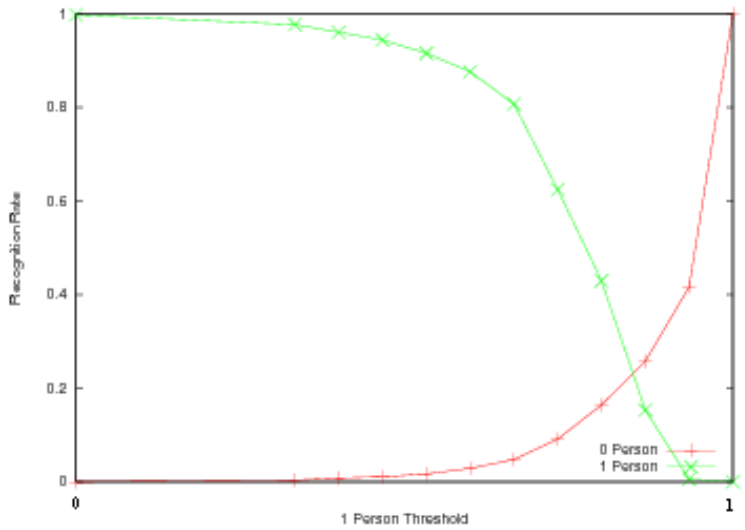


FIG. 3.15 – Résultat de classification en utilisant l'histogramme de magnitude et d'orientation du Gradient.

dernières sont statistiques. Le tableau 3.1 résume le résultat obtenu par 3 méthodes sur la même base d'apprentissage et de test.

Méthode	Personnes		Autres	
	Rappel	Précision	Rappel	Précision
Méthode structurelle basée sur crête	0.80	0.90	0.80	0.70
Méthode basée sur histogramme du Gradient	0.90	0.93	0.80	0.73
Mémoires linéaires auto-associatives	0.99	0.96	0.70	0.90

TAB. 3.1 – Comparaison de 3 méthodes de reconnaissance de personne. Les méthodes 1 et 2 n'apprennent que la classe personne. La méthode 3 nécessite d'apprendre deux classes : personne et non-personne.

Méthode 1 : Relation géométrique structurelle de quelques crêtes significatives

La représentation de personne est structurelle. Chaque personne à un instant donné est décrite par la position, la taille, l'orientation des parties du corps (ie. le torse, les jambes). Cette configuration est reconstructuable. Elle peut permettre une analyse directe de mouvement de la personne en étudiant le changement des crêtes parties.

Malgré le résultat de classification de personne et non-personne sur les séquences données soit satisfaisant et la classification s'effectue en temps réel, plusieurs améliorations sont possibles :

1. Les crêtes représentant les squelettes des structures du corps sont détectées à une seule échelle qui est l'échelle du torse. Dans un cas général, il faudrait détecter les crêtes torses à l'échelle de torse et les crêtes jambes à l'échelle de jambe.

2. L'échelle de torse et l'échelle de jambe sont déterminées en supposant que l'imagette correspond parfaitement à la personne. En réalité, la correspondance parfaite est difficile à atteindre car la segmentation reste toujours un problème primitif, une sélection automatique de l'échelle est préférable.

3. La description d'une configuration par un vecteur n'est pas discriminante. Elle ne permet pas de distinguer une imagette personne avec une imagette non-personne (ie. le cas où l'imagette contenant une crête correspondant toujours à une personne n'est pas vraie). Pour améliorer la discriminance, la structure de représentation de personne devrait être plus sophistiquée. Voir la méthode de représentation sophistiquée proposée dans le chapitre 6.

4. La représentation est spécifique pour la classe personne. Pour représenter d'autres types d'objets, une adaptation est nécessaire.

Méthode 2 : Histogramme du Gradient de la crête la plus significative

Cette méthode est une intermédiaire entre une méthode structurelle et une méthode statistique. En fait, elle utilise les crêtes pour localiser les structures importantes dans l'imagette. Pourtant, la modélisation de l'objet n'explore pas l'aspect structurel mais calcule le statistique sur l'intensité et l'orientation de changement du signal. La structure est donc représentée implicitement. Nous faisons quelques remarques suivantes :

1. L'échelle caractéristique de la composante connexe est un moyen de toutes les échelles caractéristiques à chaque point. Ce choix n'est pas correct et peut fausser le poids sur les points à l'intérieur de la crête la plus significative. Cette échelle peut être choisie autrement par exemple par segmenter les points sur la ligne de crête selon l'homogénéité de l'échelle caractéristique ou faire varier la Gaussienne le long de la crête.

2. L'histogramme de magnitude et d'orientation du Gradient n'est pas vraiment une description discriminante des objets. En fait, dans une région correspondant à une crête, il n'y a que deux directions principales opposées des points perpendiculaires à la direction de crête. Si la magnitude du Gradient est normalisée, tous les histogrammes ont des caractéristiques similaires de la forme.

3. Le fait de garder une seule composante connexe apparaît insuffisante parce qu'il y a d'autres composantes qui sont aussi significatives. Ceci revient aussi au problème classique : combien de caractéristiques sont suffisantes pour la représentation d'objet. Plus elles sont nombreuses, plus complète est la représentation mais plus coûteux est le calcul et moins tolérante à la variation.

Méthode 3 : Mémoires linéraires auto-associatives

La méthode basée sur les mémoires linéraires auto-associatives est une méthode globale. Cette méthode est simple et la plus solide en terme de programmation malgré la nécessité d'une normalisation de taille, l'orientation et l'intensité. Cette méthode comme la plupart des méthodes globale de reconnaissance est plutôt appropriée à la reconnaissance exacte mais peu adaptée à la classification.

3.8 Estimation du nombre de personne dans un groupe

L'estimation du nombre de personnes dans un groupe dans un environnement complexe est une tâche importante dans les applications de surveillance ou d'analyse de comportement. Les techniques acturelles basées sur la vision dépendent de la détection des individus. Dans le cas où les personnes dans un groupe se rapprochent visuellement, la séparation automatique des personnes reste un vrai défis.

Dans [HHD99], les auteurs ont proposé une méthode de segmentation de personnes dans un groupe en utilisant l'enveloppe convexe de la région correspondant au groupe. Ils supposent que la tête des personnes peut être trouvée sur les sommets (points de grande courbure) de cette enveloppe. Le temps pour calculer l'enveloppe est comparable par rapport à l'algorithme de tri. Pourtant, la segmentation est limitée aux personnes debout dans la scène ainsi que dans l'image. En général, les personnes peuvent être

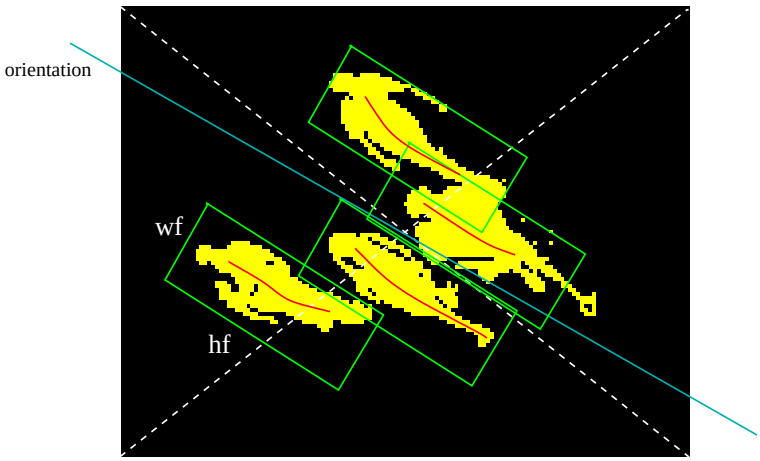


FIG. 3.16 – Segmentation de région contenant groupe de personnes

debout, assis, allongée, etc. Dans la scène, elles ne sont pas toujours verticales. Nous voulons montrer dans cette section une idée simple mais utile pour estimer le nombre de personnes dans un groupe sans aucune contrainte sur l'orientation de la personne. L'idée est basée sur l'analyse des crêtes significatives dans la région correspondant au groupe.

Premièrement, les crêtes sont détectées à quelques échelles et toutes les crêtes de longueur supérieure à un seuil sont gardées pour le test suivant. Le seuil est déterminé par la statistique sur la longueur de torse fournie par le groundtruth. Si nous connaissons la direction des personnes, nous pouvons limiter les crêtes en gardant seulement celles qui font un angle suffisamment petit avec la direction de personne (voir figure 3.16). Pour chaque crête, nous construisons le modèle de personne dans le rectangle dont le centre est le centre de gravité de la crête. La taille du rectangle relie à l'échelle et la longueur de la crête. Ce modèle est vérifié s'il est similaire à un modèle de personne appris auparavent. La détection de crêtes significatives dans la région est indépendante de la construction de modèle. Ainsi, toutes les méthodes de classification de personne et non-personne peuvent être appliquées.

La figure 3.17 montre quelques exemples de segmentation de groupe de personne. La méthode marche bien quand les personnes sont suffisamment éloignées les unes des autres. Dans d'autres cas où les personnes sont trop petites ou s'approchent trop, la détection de crête ne permet pas de détecter

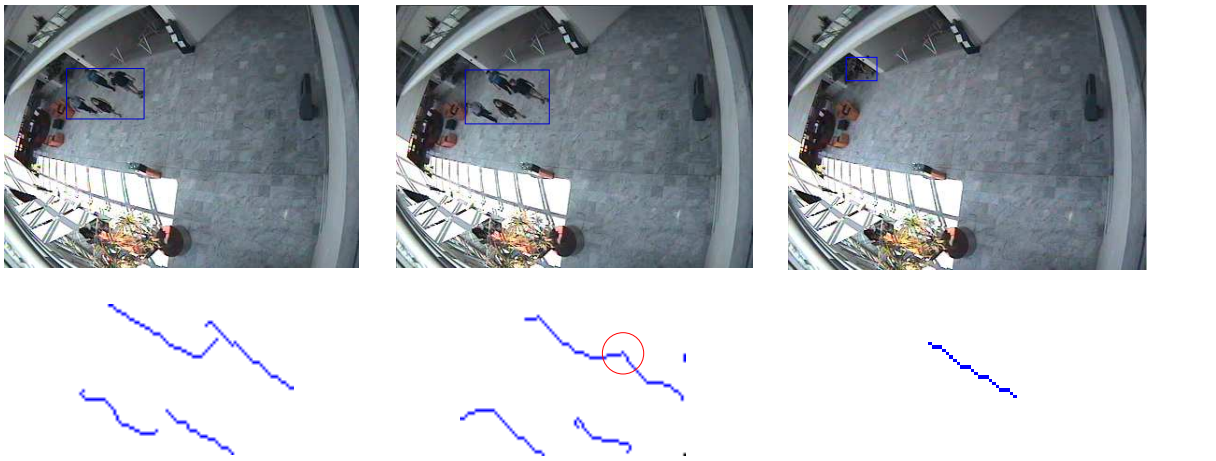


FIG. 3.17 – Exemple de segmentation de région contenant groupe de personnes

des crêtes séparées correspondant à chaque personne. La méthode 1 obtient un résultat de 51.3%. Nous espérons qu'en ajoutant les informations de contour et celles des images précédentes dans la séquence, ce résultat sera amélioré.

3.9 Conclusion

Dans ce chapitre, nous avons proposé une méthode de modélisation de personne pour la classification. La méthode est locale, structurelle. Elle permet non seulement de classifier un objet en classe personne ou non-personne mais aussi de fournir la configuration qui est utile pour l'analyse de mouvement.

33 configurations de personne ont été appris à partir de 12 séquences de vidéo. La classification a été validée sur 14 séquences réelles et un résultat de reconnaissance satisfaisant a été obtenu (rappel = 80%). Par rapport aux deux méthodes statistiques implémentées dans notre équipe, la performance est comparable (rappel = 80% pour la méthode 2 et 25% sans apprentissage de non-personne de méthode 3).

La méthode a été montrée expérimentalement robuste face au changement uniforme de la lumière. Dans le cas de changement local, la configuration de personne n'est plus correcte, mais la classification est possible. La méthode est aussi invariante à l'échelle et l'orientation parce que les crêtes sont détectées à l'échelle intrinsèque de la personne et ses angles sont calculés relativement à l'orientation de l'axe principal.

Par rapport à la méthode de modélisation de personne basée sur le squelette (l'étoile de squelettes) ou le contour, la méthode basée sur les crêtes représentent plus réellement et plus sémantiquement la configuration de la personne. Le seul problème est comment représenter la personne par des crêtes de façon plus solide et plus efficace pour mieux distinguer avec d'autres classes d'objets.

Une façon pour améliorer la discriminance de la représentation est de détecter des crêtes à leur échelle caractéristique : crête torse à l'échelle de torse, crêtes jambes à l'échelle de jambe. Il est possible

d'ajouter d'autres crêtes représentant les bras et un blob représentant la tête pour une représentation plus informative de personne. Ensuite, la description de la personne à partir de crêtes est construite de manière plus sophistiquée, par exemple une structure graphe représentant la relation de recouvrement spatial des structures.

Chapitre 4

Application à la détection de textes

Dans ce chapitre, nous présentons l'application des crêtes à la détection des textes dans des images naturelles. L'idée de base est de modéliser un texte comme une structure à plusieurs niveaux : au plus haut niveau, la ligne du texte apparait comme une crête longue, à des niveaux plus fins il y a de nombreuses crêtes courtes provenant des caractères. Ce modèle simple permet de détecter et de localiser des textes de natures très diverses dans des images, avec une précision étonnante. Cette approche est différente de celles proposées dans la littérature, qui utilisent le plus souvent l'information de couleur et de texture.

4.1 Positionnement du problème

La recherche d'images basée sur les caractéristiques telles que la forme, la couleur ou la texture n'est parfois pas satisfaisante parce que ces indices visuels ne contiennent pas explicitement la sémantique de l'image. Les textes sont fréquemment encastrés dans les images. Ils contiennent l'information importante et concise des personnes et des évènements dans la scène. L'extraction de l'information textuelle produit la sémantique de l'image, donc aide à améliorer la fiabilité de la recherche d'images.

4.2 Définition du problème

On peut distinguer un certain nombre de tâches à l'intérieur d'un système d'extraction de texte : (1) détection, (2) localisation, (3) extraction, (4) amélioration, (5) reconnaissance (OCR[1]).

La **détection de texte** consiste à déterminer la présence de textes dans l'image. Cette étape répond "non" ou "oui" pour une image donnée. Dans ce dernier cas, l'étape de **localisation de texte** construit des boites entourant les régions de textes. La détermination d'existence de texte est parfois indispensable (ie. dans les séquences vidéos où le nombre d'images contenant du texte est en général faible), parfois inutile (par exemple dans le cas des images de couvertures de livres ou d'enveloppes postales). Pour alimenter l'entrée d'un système OCR, les régions de texte qui contiennent à la fois les traits de caractères et le fond, doivent être binarisées. L'**extraction** consiste à segmenter les caractères du fond. Une phase d'**amélioration** d'image est parfois utilisée pour améliorer la qualité de la segmentation.

[1] Optical Character Recognition

Dans la littérature, les étapes “détection”, “localisation”, “extraction” sont souvent confondues. Dans ce travail, nous nous intéressons seulement à **la détection et la localisation de texte**. L'intérêt de notre modèle structurel à base de crêtes réside justement dans la qualité de la détection dans des circonstances extrêmement variées. Ce résultat peut être utilisé dans des nombreuses applications, comme par exemple l'indexation de vidéos [Che03, Wol03], l'analyse de couvertures de livres [MM99, SBK99], la détection de logos dans des émissions de télevision [PHRC03], ou dans un système de reconnaissance de caractères.

Dans le présent chapitre, nous utilisons le terme “**détection**” dans le sens “**détection et localisation**”.

4.3 Caractéristiques de textes

L'homme peut identifier rapidement une région de texte sans même reconnaître des caractères : un texte possède des caractéristiques spécifiques qui le distinguent dans une image. Intuitivement, un texte est un “alignement de caractères”, les caractères étant des lettres, ou des symboles d'un jeu de signes (que l'on n'a pas besoin de connaître dans le détail). Dans l'image, un texte est une région de forme allongée contenant un nombre important de traits. La taille, l'orientation, l'espacement des traits varient fortement selon le type de texte et l'alphabet.

4.3.1 Types de texte

Nous distinguons deux types de texte : **texte de scène** et **texte incrusté**. Un texte de scène est un texte qui a participé à la transformation scène - image lors de la prise de vue, par exemple les logos dans une émission d'un match de foot, le numéro sur le T-shirt des joueurs, les panneaux de route. Ce type de texte n'a pas reçu beaucoup d'attention de la part des chercheurs. Pourtant, il existe des applications importantes de la détection des textes de scène, par exemple la détection des plaques d'immatriculation de voitures [aQH97, SMX04], l'identification des logos [PHRC03], la détection de blocs d'adresse [WCC04], etc.

La détection de textes naturels est un sujet important dans le domaine du traitement de documents multimédia, avec des images qui rendent la détection de texte difficile : scène d'extérieur, textes superposés sur des dessins, fond arbitraire ; couleurs non uniformes. Les textes peuvent être déformés à cause de la projection perspective, de la transformation affine, du mouvement, ou de l'occlusion.

Un texte incrusté, bien au contraire, est un texte inséré après l'acquisition de l'image, avec le but bien précis de compléter l'information donnée par l'image elle-même. La police et la taille d'un tel texte sont souvent fixes. Les lignes d'un texte incrusté sont généralement horizontales ou verticales. Avec ces caractéristiques particulière, les textes incrustés sont plus faciles à détecter que les textes de scène. Le tableau 4.1 compare les caractéristiques d'un texte de scène et d'un texte incrusté.

4.3.2 Jeux de caractères

Il y a deux grands types d'écritures : alphabetique, comme l'alphabet latin, l'alphabet arabe, ou à idéogrammes, comme les texte chinois, japonais, coréen, …. Dans un texte alphabétique, les caractères sont en général de même taille et espacés régulièrement. Il en va différemment dans un texte avec des idéogrammes qui sont caractérisés différemment. La structure d'un idéogramme est plus complexe que celle d'un caractère alphabétique : un idéogramme est normalement composé de plusieurs traits de taille

Texte incrusté	Texte de scène
texte complet	occlusion partielle
indépendance de la lumière	répercutées par la lumière
horizontal/vertical	orientation aléatoire
police fixe, espacement régulier	variété dans la police, la fonte, le style. entre caractères
vue de face	projection perspective

TAB. 4.1 – Comparaison de caractéristiques de texte de scène et texte incrusté.

et d'orientation différentes. Ensuite, la taille minimale d'un caractère latin lisible est 7-8 pixels tandisqu'il faut le double pour un idéogramme. La détection de texte doit s'adapter à ces différences.

En réalité, il y a très peu de systèmes qui détectent en même temps des textes avec des lettres d'un alphabet et des idéogrammes. Dans [SKC02], des textes écrits en anglais, coréen, chinois ont été testés dans un même système de détection qui repose sur deux caractéristiques représentatives de textes : la fréquence de pixels de contour le long de lignes verticales ou horizontales et la fréquence fondamentale calculée par une transformation de Fourier. La densité de pixels de contour est une caractéristique largement utilisée pour la détection de textes (voir la section avec l'état de l'art). La fréquence fondamentale sert à réduire le taux de fausses alarmes. Cependant, cette caractéristique ne réagit pas de la même façon dans le cas de l'alphabet latin et des idéogrammes. L'expérimentation a montré que la fréquence fondamentale permet une meilleure caractérisation des idéogrammes que des textes en alphabet latin.

4.4 Méthodes existantes pour la détection de texte

Selon les caractéristiques utilisées, les méthodes de détection de texte peuvent être divisées en deux catégories : les méthodes basées sur la région et les méthodes basées sur la texture. Il existe également quelques méthodes qui utilisent ces deux caractéristiques. Un résumé des méthodes de détection de texte est présentée dans [KJ04].

4.4.1 Méthodes à base de régions

Les méthodes à base de régions utilisent les propriétés de couleur ou de niveau de gris des pixels dans une région, ou la différence par rapport aux propriétés des pixels du fond. Ces méthodes peuvent être classées en deux sous-catégories : (1) Méthodes basées sur les composantes connexes (2) Méthodes basées sur le contour. Ces deux types de méthodes fonctionnent de façon ascendante, par fusions successives de régions élémentaires, jusqu'à obtenir des rectangles entourant les régions de texte.

Méthodes basées sur les composantes connexes

Ohya *et al.* [OSA94] ont proposé de segmenter l'image par seuillage adaptatif. L'idée est de diviser l'image originale en plusieurs blocs. Pour chaque bloc, un seuil de niveau de gris est déterminé. Le seuillage se fait pour chaque pixel dans un bloc avec un seuil qui est calculé par une interpolation linéaire des seuils des pixels au centre de chaque bloc. A partir de l'image segmentée, une région est identifiée

comme une région de texte si la différence de niveau de gris de cette région et celui de rectangle englobant est suffisamment grande. L'expérimentation est réalisée sur 100 images de scène dans lesquelles les caractères sont sur les signes de route, les plaques d'immatriculation de voitures ou les panneaux des magasins. D'après les auteurs, un taux de localisation de texte d'ordre 85.5% a été obtenu. Pourtant, la méthode est limitée aux images simples. Dans le cas d'image complexe, un simple seuillage ne peut pas donner le résultat de segmentation satisfaisante.

Lienhart et Stuber [LS96, Lie96] ont utilisé un algorithme de segmentation "split-merge" plus sophistiqué que le seuillage pour classifier les pixels dans l'image en deux classes : texte et non-texte. L'algorithme de segmentation est basé sur la décomposition hiérarchique de l'image. Le processus "split"commence avec l'image entière comme un segment initial, qui est ensuite divisé en 4. La division continue pour chaque segment obtenu et s'arrête lorsque le critère de homogénéité est satisfait. La homogénéité est définie par la différence entre la valeur d'intensité la plus grande et la plus petite dans ce segment. Dans l'étape de fusion, deux segments adjacents sont fusionnés si la valeur d'intensité moyenne des deux segments est similaire. L'amélioration du résultat de segmentation est ensuite effectuée par l'analyse des contrastes pour l'image simple ou encore par l'analyse des mouvements des textes dans les images consécutives en cas de séquence vidéo. L'analyse géométrique est appliquée à la fin pour filtrer les composantes non-textuelles. L'utilisation d'une approche de segmentation d'image par split-merge est classique. La contribution principale dans cette approche est une analyse de mouvement des textes via des images consécutives dans une séquence ce qui permet d'augmenter la performance de la méthode. La méthode obtient un taux de détection intéressant : 86%-100% dans le cas d'images simples et 97%-100% dans le cas de séquences vidéo. La méthode n'est pas limitées aux textes horizontaux mais elle est très sensible aux seuils.

La méthode de Messelodi et Modena [MM99] comporte trois étapes : (i) Extraction d'objets élémentaires, (ii) Filtrage d'objets, (iii) Sélection de lignes de texte. D'abord, le pré-traitement tel que la réduction de bruit, l'amélioration de contraste, la quantisation est réalisé. Après le pré-traitement, sont effectuées la normalisation d'intensité, la binarisation et la généralisation de composantes connexes. En supposant que l'ensemble de composantes connexes sont des régions textes et non-textes, un filtrage est appliqué pour filtrer les régions non-texte. Cela est fait en analysant les propriétés géométriques des composantes connexes par exemple la superficie de la composante, la taille relative, la proximité à l'image, l'élongation, la densité, la contraste, etc. La sélection de lignes de texte consiste à grouper les composantes connexes restant de la phase de filtrage. Ceci est réalisé en utilisant des caractéristiques externes des composantes tels que la proximité, l'alignement et la comparabilité des hauteurs des caractères dans un texte. La méthode a été testée avec 100 images de couvre-livres et un taux de localisation de 91.2% a été obtenu. Une contribution de la méthode est son indépendance à l'angle de la ligne de texte. Pourtant, elle doit faire face à deux problèmes : (1) La phase de filtrage utilise un nombre important de seuils ce qui fragilise la détection, (2) Malgré la possibilité d'estimer l'angle de lignes de texte, cette méthode est limitée à des textes avec lignes droites.

Méthodes basées sur le contour

Les méthodes de détection de texte basées le contour considèrent qu'il y a un grand contraste entre les pixels de textes et ceux de fond, ce qui est caractérisé par la présence de points de contour. Ainsi, elles distinguent les régions de texte avec le font en vérifiant le nombre de contours présents dans ces régions. Plus précisément, un opérateur de détection de contour est appliqué sur l'image originale. En-

suite, une étape de groupement des contours proches pour générer les régions en utilisant un opérateur morphologique est appliquée.

Smith et Kanade [SK95] définissent une région de texte comme une structure rectangulaire horizontale qui contient un nombre important de contours. D'abord, ils appliquent un filtre différentiel de taille 3x3 et seuillent l'image résultat avec un seuil approprié pour extraire des contours verticaux. L'image obtenue est ensuite lissée pour enlever les segments étrangers ou relier des contours adjacents. Des critères géométriques sont enfin utilisés pour filtrer des régions non-textuelles. Cette méthode est limitée aux textes horizontaux dont la taille varie dans un petit intervalle.

Hasan et Karam [HK00] proposent de convertir l'image de couleur en un canal d'intensité Y avec une proportion spécifique ($Y = 0.299R + 0.587G + 0.114B$). Les contours sont ensuite identifiés par un opérateur morphologique de type gradient. En utilisant une dilatation, les contours sont groupés selon le critère de proximité pour former des régions. Le critère géométrique sur les régions elle-mêmes est là pour filtrer des régions non-textuelles. Cette approche est simple, indépendante de l'orientation du texte et de l'alignement des pixels dans une ligne de texte (possibilité de détecter des textes courbes). Pourtant, la performance de la méthode n'est pas explicitement montrée.

Une approche similaire à [HK00] est proposée par Chen *et al.* [Che03]. Les auteurs détectent les contours dans deux directions : horizontale et verticale par un détecteur de Canny. Un opérateur de "dilatation" (une structure rectangulaire de taille 5x1 et 3x6) est appliqué sur les points de contour pour générer des régions candidates. Une heuristique géométrique est enfin utilisée pour enlever des régions non-texte. Cette approche est simple, mais seulement deux directions de textes sont considérées. La méthode proposée est plutôt appliquée aux textes incrustés pour l'indexation de vidéos.

Une méthode indépendante du type d'écriture (alphabet, idéogrammes) est décrite dans [SKC02]. Cette méthode est composée de deux étapes : (1) Détection des lignes horizontales et (2) Détermination des blocs sur ces lignes correspondant à des régions de textes. La détection des lignes horizontales est réalisée en comptant le nombre de pixels ayant une valeur forte du gradient selon la direction horizontale.

Ces lignes horizontales sont supposées contenir des textes quelque part. Une étape de détection verticale permet de localiser les textes sur ces lignes. Cette détection calcule le nombre de contours significatifs, horizontaux ou verticaux dans chaque bloc de taille 8x8. En déplaçant d'un pixel à chaque pas ce bloc le long de la ligne horizontale, ils identifient un bloc comme étant une région de texte si le nombre de contours dans ce bloc est suffisamment grand. Un test d'auto-corrélation du spectre calculé sur chaque bloc est finalement utilisé pour décider si le bloc contient du texte ou pas. Cette méthode, d'après leurs auteurs, est indépendante de l'alphabet, mais est strictement limitée aux textes horizontaux et ne permet pas de variation de la taille des textes.

4.4.2 Méthodes basées sur la texture

Les méthodes basées sur la texture partent de l'observation qu'un texte a des propriétés de texture qui le distinguent du fond. Ces propriétés sont caractérisées par des techniques basées sur le filtre de Gabor, la transformation de Fourier, la variance spatiale, les ondelettes, etc. Les pixels sont classifiés grâce à un algorithme de clustering qui génère des composantes connexes. Un test géométrique est réalisé sur ces composantes pour filtrer des régions non-texte.

Wu *et al.* [WMR97, WMR99] calculent pour chaque pixel un vecteur de caractéristiques qui représentent les propriétés texturelles des pixels. Il s'agit de 9 dérivées d'ordre 2 à trois niveaux d'échelle $\sigma = 1, \sqrt{2}, 2$. Ces vecteurs sont l'entrée d'un algorithme de clustering pour classifier les pixels en trois

classes : texte, fond et non-texte. Ensuite, un processus en 5 étapes permet de déterminer les régions de texte dans l'image à chaque échelle : (1) génération de traits, (2) filtrage des traits, (3) agrégation des traits, (4) filtrage des régions, (5) extension des régions. Une phase de fusion des régions de textes à différentes échelles est réalisée à la fin. Cette approche est multi-échelle mais le fait de détecter les textes à trois niveaux seulement semble insuffisant pour une forte variation de taille des caractères dans l'image. Quelques textes de petite taille sont manqués. De plus, le calcul de 9 dérivées de Gaussien et le clustering par l'algorithme K-Means coûtent cher et ne permettent pas toujours de focaliser les régions d'intérêt. En fait, la classification des pixels en classes texte, non-texte ou fond basée sur l'observation de la variance de l'intensité des pixels est très intuitive et nécessite d'être remplacée par une phase d'apprentissage.

Au lieu de caractériser les propriétés texturelles sur chaque pixel, Clark et Mirmehdi [CM00a, CM00b] utilisent 5 mesures statistiques qui tiennent compte des voisinages autour du pixel considéré : l'information locale à chaque pixel, la densité de contour, la densité spatiale, le niveau d'asymétrie des angles des pixels voisins, l'étendue de la magnitude du Gradient selon toutes les directions. Un vecteur de 5 composantes est l'entrée d'un réseau de neurones à trois couches. Le réseau classifie chaque pixel en deux classes : texte ou non-texte. Cette approche a trois points intéressants : (1) l'apprentissage de réseaux de neurones évite l'utilisation de seuils, (2) l'indépendance de l'orientation de la ligne de texte, (3) la possibilité de détecter les textes de différentes tailles. Pourtant, il existe encore des problèmes. D'abord, l'utilisation des paramètres de la taille des masques pour calculer des mesures ne permet pas de traiter une grande variété de textes. Ensuite, l'apprentissage de réseau de neuronnes dépend fortement des exemples donnés, cequi est la faiblesse commune de toutes les méthodes basées sur l'apprentissage. En fin, aucune évaluation de performance de la méthode n'a été réalisée. Les expérimentations sont faites sur les posters dans lesquels les textes sont sur fond blanc et bien séparés d'autres objets de l'image par des bordures.

Mao *et al.* [MCLS02] ont proposé une approche qui est applicable dans le cas de texte latin et des idéogrammes. Cette méthode est réalisée à plusieurs échelles. A chaque échelle, ils calculent la variation d'énergie locale à chaque pixel. Cette variation est mesurée par une transformation d'ondelette (ie. décomposition de Harr) et seuillée par un seuil quelconque. L'image binaire obtenue est passée à la phase d'analyse de composantes connexes. Ces composantes sont finalement raffinées par une projection profil et un filtrage géométrique est utilisé pour classifier les régions de texte ou non-texte. Un avantage de l'approche proposé est son indépendance à la alphabet (testé avec l'anglais et le chinois), pourtant elle est limitée aux textes horizontaux ou verticaux à cause de l'utilisation de projection profil dans la phase d'analyse de composantes connexes. Cette approche peut être capable de détecter des textes avec une variété forte de taille si elle est réalisée à toutes les échelles dans la décomposition de Harr.

Apprentissage des propriétés textuelles

Comme l'utilisation de texture est sensible au bruit, à la taille et au style (la police) des caractères, le fait de générer manuellement un ensemble de filtres de texture adaptés à toutes les situations est impossible. Par conséquent, des méthodes d'apprentissage ont été proposées pour permettre de créer les filtres de façon automatique [JB92, JK96, Jun01, JJKK99]. Les méthodes basées sur la texture sont capables de détecter les textes sur fond complexe. Pourtant, ces méthodes nécessitent un temps de calcul de textures important. Dans le cas où le texte occupe une petite région dans l'image, le balayage de l'image entière est inutile. En plus, malgré le fait que l'apprentissage rende la détection plus robuste au changement de taille et de style de caractères, la performance de la détection dépend strictement des

Algorithme	Nature des Images	Hypothèses	Rappel-Précision
Ohya *et al.* [OSA94]	Panneaux de signalisation	Horizontal	85.5%- ?
Lienhart *et al.* [LS96]	Titre des filmes	Horizontal	86-96% - ?
Messelodi *et al.* [MM99]	Couvre-livres	Horizontal	91.2%-54%
Smith *et al.* [SK95]	Vidéo télévisée	Horizontal	93.75%-81%
Chen *et al.* [Che03]	Vidéo télévisée	Horizontal	94.51%-98.27%
Sin *et al.* [SKC02]	Vidéo télévisée	Horizontal, taille limitée	78.6%- ?
Wu [WMR97]	Vidéo télévisée	Horizontal	89.1%- ?
Mao *et al.* [MCLS02]	Images scannées	Horizontal	91.2%-88.76%

TAB. 4.2 – Comparaison d'algorithmes de détection de textes.

exemples fournis pour apprendre le système.

4.4.3 Bilan des problèmes

Nous venons de voir différentes méthodes pour la détection de textes. Nous identifions les problèmes suivants :

- *Orientation du texte* : La plupart des méthodes proposées supposent que l'orientation des textes est horizontale ou verticale. Ainsi, ils proposent des critères stricts pour détecter des textes seulement dans ces deux directions. Ces approches sont applicables aux cas des textes incrustés où 95% des textes sont horizontaux, mais se prêtent mal à la détection de textes de scène.
- *Taille du texte* : Même s'il existe quelques méthodes multi-échelle, le nombre maximal de niveaux est égal à 3, ce qui est insuffisant pour la forte variation de la taille de texte.
- *Fusion d'échelles* : Toutes les méthodes multi-échelle réalisent la détection de texte séparément à chaque niveau. Elles ont toutes besoin d'une étape de fusion des régions entre les échelles. La fusion peut prendre du temps si les échelles sont nombreuses et si l'espacement des échelles est petit, ce qui peut provoquer un grand nombre de régions qui se recouvrent (scale-redundant régions).
- *Transformations* : Les méthodes proposées détectent les textes qui sont bien formatés et bien alignés (les textes incrustés). Or dans les images d'extérieur où les textes peuvent subir n'importe quelle transformation (ie. la transformation affine), les lignes de textes peuvent avoir une forme courbe. De telles situations rendent difficile les approches existantes.

Notre objectif est de construire un détecteur automatique qui soit :

- *Capable* de détecter les textes de différentes tailles et orientations dans une image.
- *Indépendant* de la police de caractères ainsi que du système d'écriture (alphabet et idéogramme).
- *Robuste* à la déformation que subissent les caractères avec la projection perspective.
- *Générique* à tous les types de texte : texte de scène ou texte incrusté.

4.5 Méthode basée sur les caractéristiques structurelles

Un texte est un objet structuré *particulier*. A petite échelle, nous pouvons voir les traits des caractères. A une résolution plus basse, les caractères s'estompent et la ligne de texte apparaît comme une bande

(a) (b) (c)

FIG. 4.1 – (a) Exemple d'une image de texte (175x40 pixels). (b, c) Laplacien calculé à l'échelle $\sigma = 4$ et $\sigma = 16$ respectivement (valeurs positives en jaunes, valeurs négatives en vert

.

(a) (b)

FIG. 4.2 – Superposition des crêtes détectées à l'échelle $\sigma = 4$ et $\sigma = 16$ sur l'image originale : (a) à l'échelle $\sigma = 4$, les crêtes courtes correspondent aux traits des caractères. (b) à l'échelle $\sigma = 16$, une crête longue horizontale représente la ligne de texte.

allongée assez uniforme en couleur (figure 4.1). Ces structures peuvent être représentées par des crêtes à plusieurs échelles : à une échelle petite, les crêtes courtes représentent les squelettes des caractères, à une échelle plus grande, une crête plus longue représente l'axe médian de la ligne de texte (figure4.2).

La méthode de détection de texte proposée comporte deux étapes : (i) La première étape calcule les crêtes à plusieurs échelles ; (ii) La deuxième étape sélectionne les régions de texte en vérifiant des contraintes sur les longueurs et sur l'orientation des crêtes squelettes.

Ces étapes sont détaillées dans les paragraphes suivants.

Les figures 4.3 et 4.4 montrent quelques résultats obtenus avec cette méthode ; d'une façon très qualitative, on constate que les résultats sont bons dans une grande variété de situations.

4.5.1 Détection de crêtes

La détection des crêtes se fait selon les méthodes développées dans le chapitre 3.

Notons que les largeurs des textes et des caractères sont contraintes. La largeur minimale pour qu'un texte soit lisible est de 6-8 pixels et la largeur de la plupart des textes ne dépasse pas la moitié de l'image. Ce type de connaissance permet de limiter le nombre d'échelles calculées.

Soient wt_{min} et wt_{max} la largeur minimale et la largeur maximale d'une région de texte. Alors, les niveaux d'échelle des lignes de texte se trouvent dans l'intervalle $[f(wt_{min}), f(wt_{max})]$ avec $f(x) = 2(log_2(x) - 1)$. Dans le même , l'intervalle des niveaux d'échelles des traits est $[f(wc_{min}), f(wc_{max})]$ où wc_{min} et wc_{max} sont la largeur minimale et la largeur maximale des traits des caractères.

4.5.2 Vérification des régions de textes

Nous appelons des crêtes détectées au niveau $k \in [f(wt_{min}), f(wt_{max})]$ les *crêtes centrales*, les crêtes au niveau $m \in [f(wc_{min}), f(wc_{max})]$ les *crêtes squelettes*. L'identification des régions dans

FIG. 4.3 – Exemples de textes, et résultats de notre méthode : crête centrale et crêtes squelettes.

(a) (b)

FIG. 4.4 – Un exemple de texte circulaire, crêtes détectées

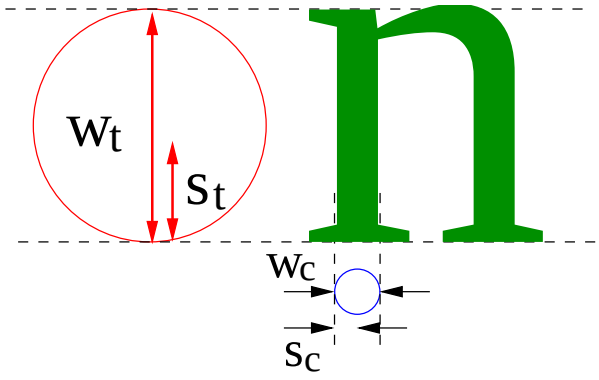


FIG. 4.5 – La hauteur d'un texte est deux fois l'échelle de la crête centrale : $w_t = 2s_t$. La largeur d'un trait de caractère est deux fois l'échelle de crête squelette : $w_c = 2s_c$.

l'image en texte ou non-texte est réalisée de la manière suivante :

- Pour chaque crête centrale T^k au niveau $k \in [f(wt_{min}), f(wt_{max})]$ correspondant à l'échelle $s_t = \sqrt{2}^k$, nous calculons la région R^k correspondant à cette crête. La région R^k est composée des points à l'intérieur des cercles de rayon s_t centré aux points sur la ligne de crête (voir figure 4.6). La région calculée de cette manière couvre la structure dont l'axe médian est la crête.

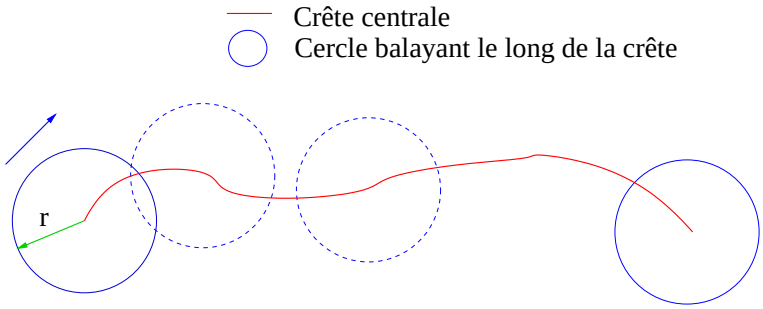


FIG. 4.6 – Construction de région associée à une ligne de crête

- La région R^k est une région de texte si les critères suivants sont satisfaits :

  CR1 : Une région de texte doit satisfaire certains critères géométriques. La plupart des méthodes existantes utilisent le rapport de la longueur et de la largeur du rectangle encadrant cette région. Ces mesures s'appliquent seulement aux lignes droites. Comme dans notre cas une ligne de texte peut être courbe, nous proposons d'utiliser l'excentricité, calculée par :

  $$e_r = 1 + \frac{l_c}{2s_t} \tag{4.1}$$

  où l_c est la longueur de la crête.

  CR2 : Avec les caractères de l'alphabet latin, les crêtes squelettes sont souvent perpendiculaires à la crête centrale au point de contact. Cette propriété n'est plus vraie dans le cas des caractères

idéogrammes (ie. le chinois, le japonais). Nous affaiblissons la contrainte de perpendicularité par la contrainte de non-parallélisme des crêtes squelettes par rapport à la crête centrale.

Le critère CR2 se base sur l'observation que chaque trait dans un texte contribue au moins une crête squelette. Ainsi la région correspondant à une ligne de texte composée de plusieurs traits doit contenir un nombre important de crêtes squelettes non-parallèles. Plus concrètement, au niveau m avec $m \in [f(wc_{min}), f(wc_{max})]$ correspondant à l'échelle $s_c = \sqrt{2}^m$, le nombre n_{sq} de crêtes squelettes C^m dans la région R^k, ayant une certaine longueur et étant non parallèles à la crête centrale T^k doit être supérieur à $nbCaracteres$, c-a-d $n_{sq} \geq nbCaracteres$ où

$$n_{sq} = \#\{C^m \in R^k; lc_{min} \leq l_{C^m} \leq lc_{max}, C^m \not\parallel T^k\} \tag{4.2}$$

- Lorsqu'une région est identifiée comme région de texte, cette région n'est plus considérée par la suite. De cette façon, l'algorithme de détection ne doit pas vérifier toutes les crêtes centrales aux tous les niveaux. Par conséquent, le temps de calcul est considérablement réduit.

La vérification du critère **CR2** nécessite de répondre au quatre questions suivantes :

1. ***Quel est le niveau d'échelle*** m ***des crêtes squelettes***, étant donné le niveau de la crête centrale k et sachant que $m < k$ et $m \in [f(wc_{min}), f(wc_{max})]$? La solution la plus simple est de vérifier le critère **CR2** à tous les niveaux m possibles dans l'ordre décroissant. Dès qu'à un niveau m, la région est identifiée comme région de texte, la vérification est arrêtée pour cette région.

   Le niveau d'échelle des crêtes squelettes trouvées de cette manière peut ne pas correspondre à l'échelle optimale des traits de caractères parce des crêtes représentant une même structure se répètent souvent à plusieurs échelles. Une solution est de chercher parmi les niveaux dans l'intervalle défini un niveau auquel la déviation de l'échelle s_c par rapport à l'échelle optimale de chaque point sur les crêtes squelettes est la plus petite.

$$dev = \sqrt{\frac{1}{\#P_i} \sum_{P_i \in C^m} (s_{optimal} - s_c)^2} \tag{4.3}$$

2. ***Comment déterminer si une crête*** C^m ***au niveau*** m ***peut être une crête squelette de la région*** R^k ? Pour cela, deux tests sont proposés : le test d'appartenance et le test de longueur.

   Le test d'appartenance d'une crête C^m à la région R^k est réalisée en calculant le nombre de points de la crête C^m à l'intérieur de cette région. Une appartenance entière est obtenue si tous les points de la crête sont dans la région. Comme une crête peut être fusionnée avec une autre crête à cause du lissage, nous tolérons ce cas en acceptant des crêtes qui ont 80% des points dans la région.

   Le test de longueur suppose que la crête correspondant au squelette d'un trait ne doit pas être trop courte ni trop longue. Elles sont limitées dans un intervalle $[lc_{min}, lc_{max}]$. Dans le cas où l'on ne connait pas lc_{min} et lc_{max} a priori, ces valeurs sont estimées selon le raisonnement suivant : D'abord, une crête squelette détectée à l'échelle s_c devrait avoir une longueur supérieure à deux fois celle de l'échelle, c-à-d $2s_c$ (sinon, la crête va s'allonger dans la direction perperdiculaire). Comme le lissage racourcit la crête correspondant à un trait rectangulaire (aux deux extremités), nous posons $lc_{min} = s_c$. Ensuite, le rectangle encadrant une crête squelette devrait avoir une hauteur inférieure à la hauteur de la ligne de texte. Le lissage peut rassembler les structures et les rendre plus longues. Ainsi nous multiplions la valeur estimée par un facteur 3, donc $lc_{max} = 3s_c$.

3. ***Comment vérifier une relation non-parallèle ?*** Une relation non-parallèle est mesurée par l'angle entre deux crêtes. Comme une crête n'est pas forcément une ligne droite, l'angle entre deux crêtes est difficile à déterminer. Normalement, la crête centrale et les crêtes squelettes s'intersectent à peu près au centre de la crête squelette. Nous déterminons la relation non-parallèle entre deux crêtes par l'angle entre deux vecteurs principaux au point d'intersection de deux crêtes. Le vecteur principal d'un point est le vecteur correspondant à la valeur propre la plus petite de la surface associée à ce point à l'échelle considérée. La figure 4.7 montre deux directions prinpales à un point d'intersection de la crête centrale et de la crête squelette à deux niveaux différents.

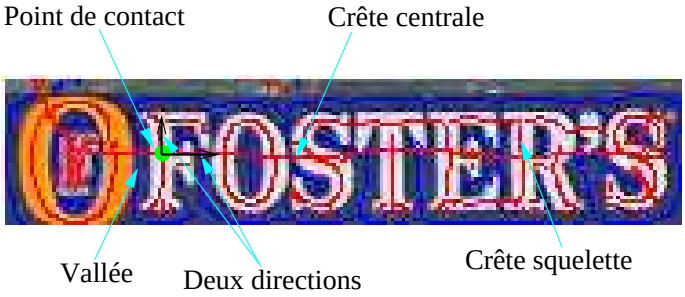


FIG. 4.7 – Crêtes détectées aux échelles $\sigma = 4$ et $\sigma = 16$ superposées sur l'image : les crêtes sont en rouge, les vallées sont en bleu.

4. ***Combien de crêtes squelettes sont suffisantes, c-à-d que vaut*** $nbCharacters$ ***?*** Dans notre expérimentation, nous choisissons $nbCharacters = 2$. Cette valeur est raisonnable parce qu'en réalité, un texte contient au moins 2 caractères. Quand il a un seul caractères, il s'agit d'un caractère isolé. Notre méthode détecte les lignes de texte. Elle n'est pas faite pour détecter des caractères isolés.

   Apprentissage automatique des directions et des longueurs de crêtes squelettes : Le critère non-parallélisme, bien qu'il soit plus faible que le critère de perpendicularité, n'est pas générique pour de tous des caractères. Pour les caractères idéogrammes, c'est toujours le cas. La figure 4.8 montre le résultat de détection de textes en vérifiant les deux critères **CR1** et **CR2**. Nous constatons qu'avec le critère non-parallélisme des crêtes squelettes par rapport à la crête centrale, seule la ligne de texte en bas contenant des traits perpendiculaires est détectée. La plus grande ligne de texte en haut contennant plusieurs traits parallèles à la ligne centrale du texte n'a pas été détectée. Dans un texte idéogramme, il arrive souvent qu'un trait soit parallèle à la direction de la ligne de texte. Ainsi il faudrait apprendre cet arrangement spatial des caractères afin de prendre la décision. L'apprentissage peut être réalisé en utilisant un réseau de neurones ou un SVM (Support Vector Machine).

4.6 Evaluation

4.6.1 Critères d'évaluation

La performance de la méthode de détection de textes est évaluée en termes de taux de rappel et de taux de précision. Ces mesures de performance sont calculées en termes de régions de texte, non en

FIG. 4.8 – Résultat de détection de texte. La région correspondant à la ligne de texte en haut ne satisfait pas le critère de non-parallélisme parce que'elle contient plusieurs traits parallèle à la ligne de texte.

termes de pixels textuels.

Le rappel est défini par le ratio entre le nombre de véritables régions de textes détectées et le nombre de régions de textes réelles existant dans l'image :

$$\text{Rappel} = \frac{\#\text{Vraies Détections}}{\#\text{Nombre réel de textes}}$$

La précision est définie par le ratio du nombre de vraies régions de textes et du nombre total de régions détectées :

$$\text{Précision} = \frac{\#\text{Vraies Détections}}{\#\text{Vraies Détections} + \#\text{Fausses Détections}}$$

4.6.2 Images de test

L'algorithme de détection de texte est testé sur les bases de 145 images de natures différentes. Le résumé de l'information de ces bases de données est montré dans le tableau 4.3. Quelques images extraites de ces bases sont montrées dans la figure 4.9.

	#Images	Résolution	Caractères	Type de texte
DB1 : Transparents	10	640x480	alphabet	texte de scène
DB2 : Journaux télévisés	45	352x288	alphabet	texte incrusté
DB3 : Sous-titres de filme	20	352x288	alphabet	texte incrusté
DB4 : Villes Asiatiques	50	352x288	idéogramme	texte de scène
DB5 : Course de voitures	20	352x288	alphabet	mixte

TAB. 4.3 – Bases de données utilisées pour tester l'algorithme de détection de texte.

Dans ces bases, la résolution maximale de l'image est 640x480. Alors, en théorie, il faut détecter les crêtes à N = $\log_2$(640x480) $= 18$ échelles. Pourtant, la taille des textes varie seulement dans l'intervalle

$[4, 73]$. La détection en pratique ne se réalise qu'à N = $\lceil 2\log_2(73/2) \rceil$ = 11 échelles. En fait, les crêtes détectées aux niveaux supérieurs à 11 représentent les structure de largeur supérieure à $2 * \sqrt{2}^{11}$ = 90 pixels donc ne sont plus structures textuelles. Le tableau 4.4 résume les valeurs de seuils que nous utilisons pour le test.

Notation	Signification	Valeur
wc_{min}	Largeur minimale de trait	2
wc_{max}	Largeur maximale de trait	16
$lc_{min} = wt_{min}$	Largeur minimale de ligne de texte	4
$lc_{max} = wt_{max}$	Largeur maximale de ligne de texte	73
nbCaracteres	Nombre minimal de caractères dans un texte	2
[sl_1,sl_2]	Intervalle de niveau d'échelle de trait	[2,8]
[tl_1,tl_2]	Intervalle de niveau d'échelle de ligne de texte	[4,11]

TAB. 4.4 – Les valeurs utilisées pour le test.

4.6.3 Résultat de détection de textes

Le tableau 4.5 montre le résultat de détection de textes dans les bases données. La détection de texte est meilleure sur les images de transparent (100%). Les images de courses de voitures sont difficiles à détectées mais le résultat est très satisfaisant (90%). Les textes idéogrammes donnent également un bon résultat (77.8%). Les sous-sections ci-dessous analysent chaque cas plus en détail.

	#Images	#Mots	#Mots détectés	#Fausses alarmes	Rappel(%)	Précision(%)
DB1	10	172	172	7	100	96.09
DB2	45	217	169	114	77.88	59.71
DB3	20	1179	934	68	79.21	93.21
DB4	50	680	529	30	77.79	94.63
DB5	20	199	177	18	88.9447	90.7692

TAB. 4.5 – Evaluation de la performance de l'algorithme de détection de texte.

DB1 : Images de transparents

Cette base contient 10 images de résolution 640x480 prises par une caméra. Il s'agit de transparents de présentation avec un fond de couleur uniforme (souvent beige ou blanc). La couleur de texte est bien différente du fond. Les textes ont des polices et des tailles variées mais la distribution des caractères est suffisamment dense pour former des lignes de texte.

La détection de texte a été réalisée correctement sur les images de transparents. Toutes les lignes de texte ont été parfaitement localisées. Les textes de différente taille sont détectés à des échelles différentes. Nous constatons que les crêtes centrales se trouvent exactement au centre des lignes de texte. Les crêtes

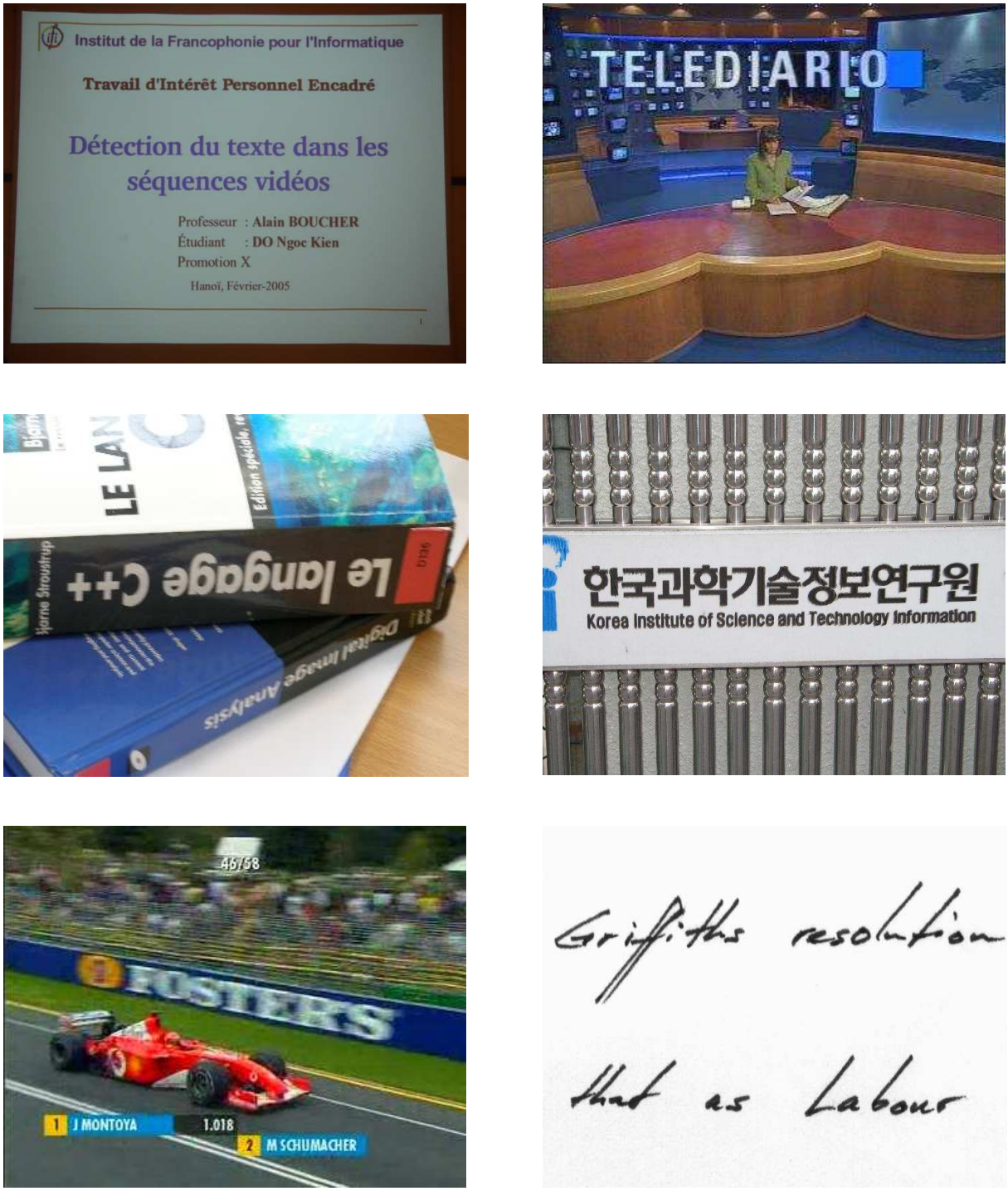
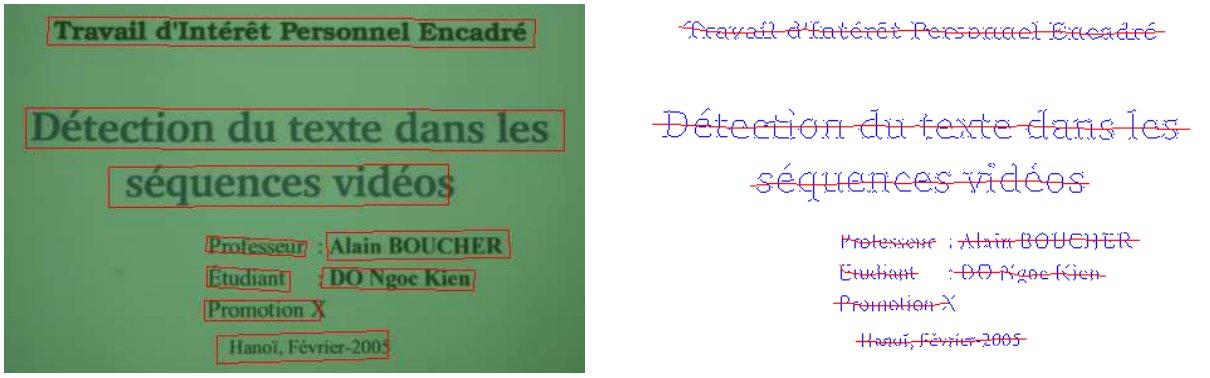


FIG. 4.9 – Quelques images des bases de données utilisées pour tester la méthode proposée.



(a) (b)

FIG. 4.10 – (a) Image d'un transparent ; Les régions de textes détectées sont encadrées par des rectangles rouges. (b) Crêtes détectées à deux échelles : $\sigma_1 = 2\sqrt{2}$ (bleu) et $\sigma_2 = 16$ (rouge) : les lignes rouges représentent les axes médians des textes, les lignes bleues représentent les squelettes des caractères.

squelettes correspondent aux squelettes des caractères, qui peuvent entrer directement dans un système de reconnaissance de caractères sans traitements supplémentaires.

Nous remarquons également que dans un document contenant des lignes de textes arrangées l'une après l'autre, entre deux lignes de textes, une crête au centre de deux lignes a été détectée. Nous appelons cette crête l'inter-ligne. Cette ligne est complètement "blanche" dans le sens qu'il n'y a aucun trait dans la région associée. Ce type de ligne détectée peut permettre de distinguer des lignes de textes ou des paragraphes.

DB2 : Images extraites de journaux télévisés

Cette base comprend 45 images extraites de journaux télévisés. Tous les textes dans ces images sont en anglais et la plupart sont de type "texte incrusté". Les textes incrustés sont tous horizontaux. La taille de texte est différente d'une image à l'autre. Dans une image, les textes ont au plus 2 tailles différentes. Les textes sont en majorité visibles sauf quelques textes portant la date des évènements qui sont petits et obscurs. Les textes incrustés sont insérés sur des fonds assez complexes et texturés. Il y a trois images sans textes. Cette base d'images a été utilisée dans [XHZ01] pour l'évaluation de la performance de leur méthode.

Intuitivement, la détection de textes incrustés est plus facile par rapport aux textes de scènes parce que les textes incrustés sont crées dans le but d'être lisibles par un être-humain. Ainsi, ils ont des caractéristiques plus particulières par rapport à un texte de scène par exemple la taille de caractères est dans un intervalle fixe, la police est simple, et surtout il est horizontal.

Nous avons utilisé un détecteur générique pour détecter des textes sans limiter la taille et l'orientation du texte. L'expérimentation a montré que la plupart des textes ont été trouvés. La localisation n'est pas parfaite parce qu'il existe des caractères majuscules et minuscules dans une ligne de texte. La différence en taille des caractères nécessite de détecter la ligne de crête centrale à des échelles caractéristiques. Notre approche détecte les points formant la crête centrale à une échelle unique. Ainsi, quand il y a une

FIG. 4.11 – Deux exemples de détection de textes incrustés dans une base d'images de journaux télévisés.

variance forte en largeur d'une ligne de texte, la localisation n'est plus exacte.

Nous trouvons que dans cette base d'images, les textes incrustés sont très souvent arrangés sous forme d'un tableau. La séparation entre deux lignes de texte n'est pas nette. Ainsi, il est parfois difficile de trouver des crêtes principales au centre des lignes de textes. Elles étaient soit déplacées légèrement à cause de l'influence de fond ou de la ligne de côté, soit manquées. Pour cette raison, certaines régions de textes n'ont pas été détectées ou quelques morceaux de texte ont été manqués.

L'algorithme a également détecté des régions qui n'étaient pas des régions de texte. Les fausses détections ont été trouvées dans des images avec un fond complexe ou texturé. Une des raisons en est que la structure "une crête principale et plusieurs crêtes ourtes" est vraie non seulement avec les régions de texte mais aussi avec des "grilles". L'élimination des fausses alarmes peut se faire en passant à l'étape de reconnaissance des caractères.

DB3 : Séquences vidéo

Nous avons pris quelques séquences vidéo de différentes natures : film, publicité, musique, etc. avec pour but qu'elles soient capables de couvrir des situations de détection de texte faciles aussi bien que difficiles. La scène et les textes dans ces vidéos sont soit mobiles soit fixes. Le mouvement de textes ou de scènes provoque l'occuration des textes au fond. Ces séquences ont été utilisée dans le travail de Lienhart *et al.* [LW02] pour tester la performance d'une approche basée sur l'apprentissage par un réseau neuronal des vecteurs de l'orientation des points de contour. Le rappel de détection sur l'image fixe par leur méthode est de 69.5% et la précision est de 47.69%. La performance du système est considérablement améliorée en tenant compte des redondances temporelles dans la séquence vidéo.

Nous réalisons l'expérimentation de notre algorithme de détection de textes basé sur des crêtes image par image dans les séquences. 80% régions de textes ont été détectées correctement. La localisation n'est pas parfaite. Plus précisément, le rectangle encadrant une région de texte est souvent plus large que la région elle-même. Cela est expliqué par le fait que nous avons discrétisé l'espace d'échelle. Quand la largeur réelle de la région de texte n'est pas égale à deux fois une des ces valeurs, cette largeur détectée

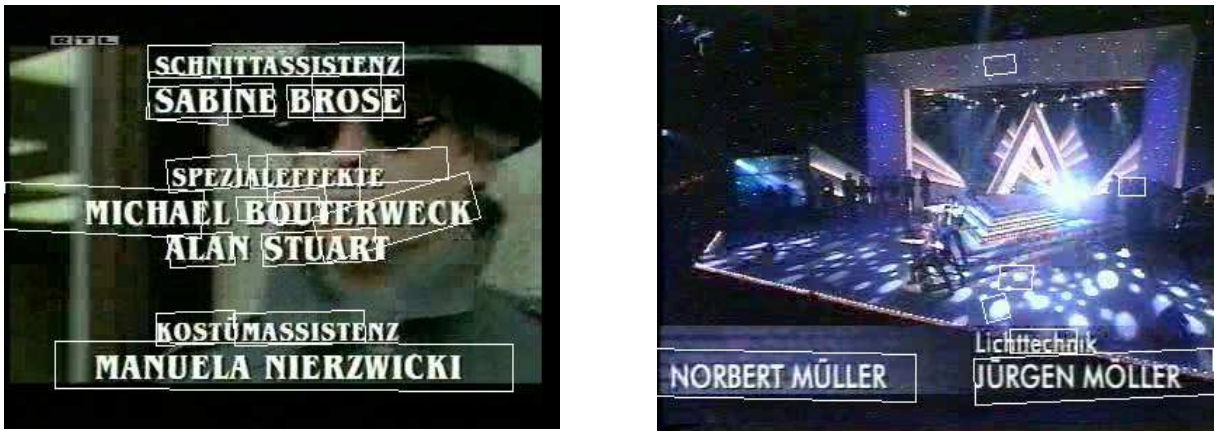


FIG. 4.12 – Détection de textes en séquences vidéo. Il y a très peu de fausses détections. La localisation n'est pas parfaite : Le rectangle encadrant une région de texte peut être plus grand que la région elle-même.

est approximative. Pourtant, l'extension du rectangle encadrant une région de texte n'est pas grave parce que nous pouvons toujours raffiner cette région par un algorithme par exemple projection de profil. Nous trouvons également que des images dans la séquence "musique" ont un fond très complexe. Pourtant, il y a très peu de fausses alarmes de la scène, ce qui montre la performance du modèle de texte basé sur les crêtes par rapport aux contours.

DB4 : Images de ville de la Corée et de la Chine

Cette base consiste en 50 images prise en Corée et en Chine. Les textes dans l'image sont les textes coréens ou chinois mélangés avec les textes en alphabet latin. Ils sont de type "texte de scène". Les images dans cette base sont souvent texturées, surtout celles prises aux pagodes en Chine. Les caractères idéogrammes sont espacés les unes des autres, ce qui rend la détection des lignes de texte difficile.

L'expérimentation montre que la représentation de texte par une crête centrale et plusieurs crêtes courtes est générique pour plusieurs types de alphabet. Quand les caractères dans une ligne de texte ont une distribution suffisamment dense, une crête longue au centre de texte a été toujours détectée. Une caractéristique des textes idéogrammes est que l'arrangement des traits n'est pas régulier. Ainsi, le critère sur le nombre de crêtes squelettes non-parallèle à la crête centrale provoque la manque de détection. Le rappel dans ce cas est 77.79% et la précision est 94.63%.

A gauche de la figure 4.13 montre un résultat parfait de détection des textes idéogrammes. Cette image est très texturée. Il y a plusieurs structures de grilles qui peuvent fausser la détection. Deux lignes de textes sont constituées à partir de caractères horizontaux ou verticaux. En dépit de ces difficultés, notre algorithme produit deux rectangles qui couvrent exactement les textes souhaités. A droite de la figure 4.13 présente la détection de texte d'une image plus simple mais la localisation de texte obtenue est moins bonne. La raison en est que la taille du texte est grande. Comme nous l'avons dit auparavant, la localisation n'est plus exacte quand la taille de texte n'est pas précisément égale à 2 fois une des valeurs d'échelle considérées. En plus, l'espacement entre deux échelles augmente exponentiellement

FIG. 4.13 – Résultat de détection de textes idéogrammes : textes chinois et coréens.

selon l'échelle. Ainsi, plus la taille de texte est grande, moins la localisation est précise.

DB5 : Images extraites des vidéos de course de voiture

Nous avons extrait 20 images dans des vidéos de courses de voitures Formule 1. Les textes dans ces images sont très difficiles à détecter. Il s'agit de textes de scène qui se trouvent sur un fond en mouvement. L'image a été prise avec une caméra mobile qui provoque une projection perspective. Par conséquence, la taille de caractères varie fortement d'un texte à l'autre et d'une image à l'autre. L'orientation des textes est aléatoire.

Malgré tous ces difficultés, notre algorithme a détecté tous les textes lisibles par l'œil humain. Il a localisé correctement les textes qu'ils sont très difficiles à cause de l'étirement et des artefacts (ie. le logo FORSTER dans la scène). Nous avons détecté des contours sur de telles images et constatons que les contours ne sont pas représentatifs pour distinguer les régions textes et non-textes. Dans la région contenant le logo FOSTER'S, il y a très peu de contours à cause d'un fort flou de mouvement. Or, dans les régions où apparaît le public, il y a de nombreux contours qui ne sont pas significatifs mais qui faussent la détection de texte par le contour.

L'expérimentation sur cette base d'images montre une indépendance à l'orientation de notre méthode. Elle détermine l'orientation automatiquement grâce à l'orientation de la crête centrale. Cette approche ne confond jamais les voitures et les textes, ce qui est commis par la méthode d'identification des textes par l'histogramme de couleur [HPRC04].

Comparaison avec le résultat de détection de textes proposé par Wolf []

Nous avons testé la détection des textes par un algorithme proposé par Wolf[2]. La figure 4.16 montre quelques résultats founis par chacun des deux algorithmes. Nous pouvons voir que notre approche est

[2]Le programme se trouve sur http ://telesun.insa-lyon.fr/%7Ewolff/demos/textdetect/html

FIG. 4.14 – Résultat de détection de textes de scènes de l'orientation aléatoire et subissant une projection perspective.

(a) (b)

FIG. 4.15 – Quelques exemples de fausse détection : (a) Structures de grilles ressemblant à un texte. (b) Le fond texturé se confond avec des textes.

capable de détecter des textes courbes d'orientation aléatoire. Les approches existant comme par exemle celle de C. Wolf détectent les caractères mais pas les bonnes lignes de textes. Dans le cas d'une image texturée, notre approche détecte une ligne de texte malgré quelques fausses détections.

4.7 Conclusion sur la détection de texte

Dans ce chapitre, nous avons présenté une nouvelle méthode pour détecter des textes dans une image. Toutes les méthodes existantes exploitent les propriétés géométriques de textes et les représentent par des vecteurs numériques qui sont absolument abstraits pour humain. Nous proposons d'analyser les textes par leur topologie et d'étudier le changement de topologie à travers les échelles. Un texte est un objet structuré de façon simple : un texte est composé des caractères qui sont considérés comme des détails de l'objet et l'ensemble de ces caractères donnent une forme longiligne au texte. Ainsi, nous avons proposé d'étudier le changement de topologie à deux échelles : une échelle grande représente la ligne de texte et une plus petite représente les traits des caractères.

Nous avons utilisé les crêtes pour caractériser les structures des textes. Un texte est modélisé par une crête longue à une échelle grande représentant la ligne centrale et plusieurs crêtes plus courtes représentant les squelettes des caractères à une échelle petite. Cette représentation a été montrée par expérimentation comme étant générique pour plusieurs types de texte, et indépendante du alphabet d'écriture, de l'orientation de texte et robuste à la transformation affine.

L'expérimentation a montré une bonne performance de notre méthode. La détection est difficile dans le cas où les lignes de textes ne sont pas bien séparées. Dans d'autres cas, la localisation est bonne. Il y a quelques fausses alarmes qui peuvent être enlevées par une étape de reconnaissance de caractères. En résumé, les contributions principales de notre travail présenté dans ce chapitre sont d'avoir proposé une approche de détection de texte qui est :

- Indépendante du jeu de caractères (alphabet).
- Capable de détecter les textes de différentes tailles.
- Indépendante de l'orientation des textes et la forme de la ligne de texte.
- Robuste aux transformations géométriques (ie. transformation affine).
- Capable de prévoir la taille des textes et des caractères grâce aux échelles.

Malgré le résultat satisfaisant de la détection sur différentes images naturelles, la méthode éprouve quelques difficultés :

- Elle ne peut pas détecter des textes dont les caractères se dispersent considérablement car dans ce cas il est difficile de détecter une crête représentant la ligne centrale de la région de texte. Il n'y a pas de solution dans ce cas parce que la méthode est faite pour la détection de lignes de texte, pas des caractères isolés.
- On ne sait pas exactement à quelle échelle il faut détecter les crêtes centrales et à quelle échelle il faut détecter les crêtes squelettes. Actuellement, la méthode proposée fait une recherche exhaustive sur l'intervalle d'échelles possibles et s'arrête lorsqu'une région est identifiée. La recherche exhaustive s'effectue à deux niveaux : au niveau du squelette et au niveau des ligne de texte.
  Une solution est de détecter des points à plusieurs échelles caractéristiques. Un point de contact d'intersection entre la ligne centrale du texte et une squelette a deux échelles caractérisriques : celle du squelette et celle de la ligne de texte. Ce point sera considéré deux fois. La détection à l'échelle caractéristique devrait faire face aux discontinuités possibles des crêtes. De plus, la

FIG. 4.16 – Comparaison avec le résultat de détection de textes proposé par C. Wolf [] : à gauche : les résultats founis par C. Wolf, à droite : les résultats founis par notre algorithme.

connexion entre les points de similaire niveau d'échelle doit être réalisée.

- Plusieurs valeurs de seuils sont utilisées. Ces valeurs peuvent être calculées automatiquement si une approche à échelle caractéristique est appliquée.

Chapitre 5

Représentation hiérachique structurelle par crêtes et pics

Dans cette thèse, nous avons étudié les crêtes en tant que caractéristiques visuelles, et nous avons montré l'utilité des crêtes dans deux applications spécifiques : la reconnaissance de personnes et la détection de texte. Ces deux études de cas sont encourageantes, car les crêtes fournissent des structures dans une image que l'on peut facilement exploiter dans des traitements plus "sémantiques" ultérieurs.

Ceci étant, nos études de cas restent très simples. Nous voulons donc les placer ici dans une perspective ambitieuse, dans le cadre d'un travail à beaucoup plus long terme : la représentation d'objets par des crêtes pour la reconnaissance générique.

Nous proposons de modéliser par des crêtes détectées à plusieurs échelles. La forme globale de l'objet est caractérisée par les crêtes à grande échelle tandis que les détails sont représentés par les crêtes aux échelles plus petites. La description gros-détail permet de représenter à la fois des objets de différentes classes (en considérant les structures grossières) et de discriminer des variantes à l'intérieur d'une classe (en considérant les détails).

Les caractéristiques grossières et détaillées sont organisées dans une structure de graphe dont la racine correspond au niveau d'abstraction de forme le plus haut et dont les feuilles correspondent au niveau le plus détaillé. Cette représentation hiérarchique permet une stratégie de mise en correspondance hiérarchique efficace, ce qui n'est pas le cas de la mise en correspondance de graphes en général.

Une première étude de cette approche a été réalisée dans un cadre d'un DEA [Pha05] dont nous donnons quelques éléments ci-après.

5.1 Représentation multi-échelle par crêtes

Cette section présente les étapes principales pour construire une représentation hiérarchique avec des crêtes, et expose les premiers résultats d'expérimentation de mise en correspondance de modèles.

5.1.1 Construction du graphe attribué relationnel

Un graphe attribué relationnel [EF86] est un graphe dont chaque nœud et chaque arc est étiqueté avec des "attributs" qui sont des informations numériques ou relationnelles. Ces informations attachées

aux nœuds et aux arcs permettent une représentation plus complète et plus discriminante qu'un graphe simple. Nous adaptons cette idée à la construction de notre graphe de représentation à base de crêtes.

Un graphe attribué relationnel est donné par T = (V, E, R, H) où V est l'ensemble des nœuds, E l'ensemble des arcs, R l'ensemble des attributs-étiquettes des nœuds et H l'ensemble des attributs-étiquettes des arcs.

1. *Nœud et attributs* : Chaque caractéristique crête ou pic définit un nœud dans le graphe de représentation. Chaque nœud possède les attributs suivants :
   - Type de caractéristiques : Crête, Vallée, Pic, Pic négatif.
   - Echelle à laquelle la caractéristique est détectée.
   - Longueur, énergie moyenne.
   - Mesures statistiques calculées sur la région de couverture de la caractéristique (voir la figure 5.1) : contexte de la forme (shape context), histogramme des directions (figure 5.2), vecteur de signature topologique (voir ci-après), ….

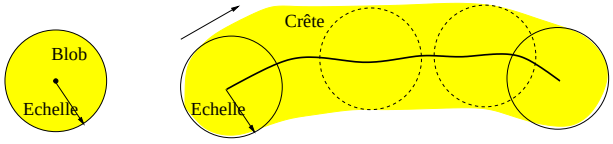


FIG. 5.1 – Région associée à une caractéristique. Pour un pic, c'est une région circulaire de rayon σ. Pour une crête, elle est déterminée par balayage d'un cercle de rayon σ le long de la crête.

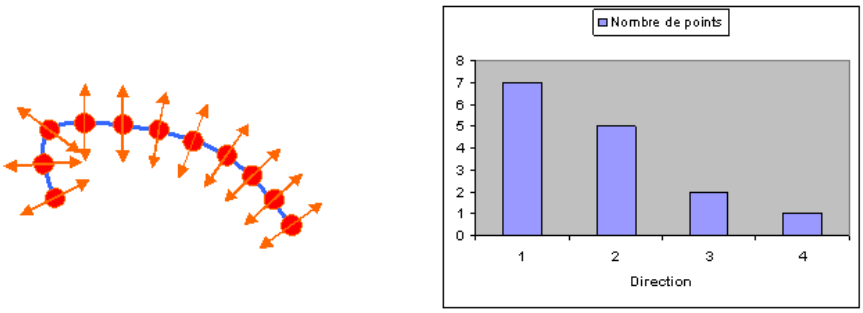


FIG. 5.2 – Histogramme des directions principales des points d'une crête.

2. *Arc et attributs*. Un nœud i à niveau k dans le graphe est dénoté par ${N_i}^k$. Un arc est établi entre deux nœuds ${N_i}^k$ et ${N_j}^{k+1}$ lorsque la région de couverture correspondant à la caractéristique du nœud ${N_j}^{k+1}$ au plus haut niveau couvre suffisamment la région de couverture de la caractéristique du nœud juste en dessous (voir la figure 5.3). Dans notre implémentation, nous calculons le rapport des points communs entre deux régions correspondantes aux deux nœuds ${N_i}^k$, ${N_j}^{k+1}$ et des points dans la région associée au nœud ${N_i}^k$. Si ce rapport est supérieur à un seuil, un arc est construit entre deux nœuds : ${N_j}^{k+1} \rightarrow {N_i}^k$.

   Les attributs d'un arc sont :
   - Type : Crête-Vallée, Vallée-Crête, Crête-Pic Négatif, Crête-Pic, etc.

- Rapport d'échelle normalisé.
- Angle entre deux directions principales.

Dans [SMD+05], Shokoufandeh *et al.* ont montré qu'il est possible de représenter la topologie d'un graphe par les valeurs propres de la matrice d'adjacence. D'où l'idée de caractériser chaque nœud du graphe par un vecteur qui est composé de k éléments déterminé de la manière suivante (k est le facteur de branchement maximal du graphe).

Pour chaque nœud, on détermine la matrice d'adjacence de son sous-graphe. On calcule les valeurs propres de cette matrice et garde la valeur de la somme des k_i valeurs propres les plus grandes, notée $S_i = \lambda_1 + \lambda_2 + ... + \lambda_{k_i}$. Pour former le vecteur de signature topologique d'un nœud V, on combine les valeurs S_i de ses fils par ordre décroissant.

5.1.2 Mise en correspondance hiérarchique

Les graphes représentant les objets d'apprentissage seront stockés comme des graphes-modèles. La reconnaissance d'un nouvel objet dans une image consiste à mettre en correspondance le graphe du nouvel objet avec les graphes appris. Ce paragraphe présente l'appariement de deux graphes de représentation.

La mise en correspondance de graphes de représentation, en tant que problème de mise en correspondance des graphes, est un problème bien connu pour sa complexité algorithmique. Dans la littérature, on s'intéresse à trois formes du problème de mise en correspondance de graphes :

- *Isomorphisme de graphes* consiste à vérifier si les structures de deux graphes sont identiques.
- *Isomorphisme de sous-graphes* consiste à chercher des isomorphismes entre le graphe 1 avec les sous-graphes du graphe 2.
- *Double isomorphisme de sous-graphes* consiste à chercher tous les isomorphismes entre les sous-graphes du graphe 1 avec les sous-graphes du graphe 2.

D'une façon générale, le problème d'isomorphisme de graphe est un problème NP-complet, ce qui implique qu'aucun algorithme efficace (de complexité non exponentielle) de solution n'est connu.

Pour cette raison, nous proposons une stratégie de mise en correspondance hiérarchique, spécifique à la structure de nos graphes attribués, et qui conduit à un algorithme efficace. Notre algorithme se base sur l'observation que si deux objets sont similaires, alors, les caractéristiques qui les représentent doivent être similaires et avoir la même organisation spatiale et hiérarchique. L'algorithme est du type "glouton" : il réalise la mise en correspondance à chaque niveau du graphe et s'arrête lorsqu'une réponse est considérée fiable.

Soit G_n le graphe de l'objet à reconnaître, G_i le $i$ème graphe modèle. La mise en correspondance se compose de 3 étapes :

- Etape 1 : Comme les objets peuvent subir un changement d'échelle, la première étape est la recherche d'un nœud V dans l'graphe G_n qui ressemble le plus à la racine du graphe modèle G_i. La mise en correspondance n'est réalisé qu'entre le sous-graphe du nœud V du graphe G_n avec G_i. Nous appelons ce sous-graphe G_{sn}.
- Etape 2 : Pour chaque niveau suivant, on vérifie les contraintes symboliques des caractéristiques des arcs et mesure la similarité des nœuds pour déterminer la paire des nœuds correspondante.
- Etape 3 : Itérer l'étape 2 jusqu'à ce que l'on ne trouve plus de nœud dans le prochain niveau ou que les graphes ont atteint à leur hauteur.

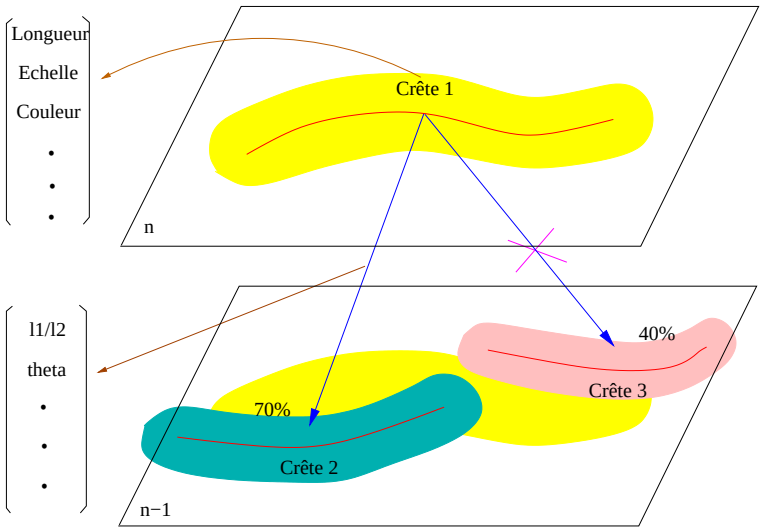


FIG. 5.3 – Construction des nœuds et des liens : Chaque crête/pic définit un nœud. Le pourcentage de recouvrement spatial de deux régions associées à deux caractéristiques détermine un arc ou l'absence d'arc.

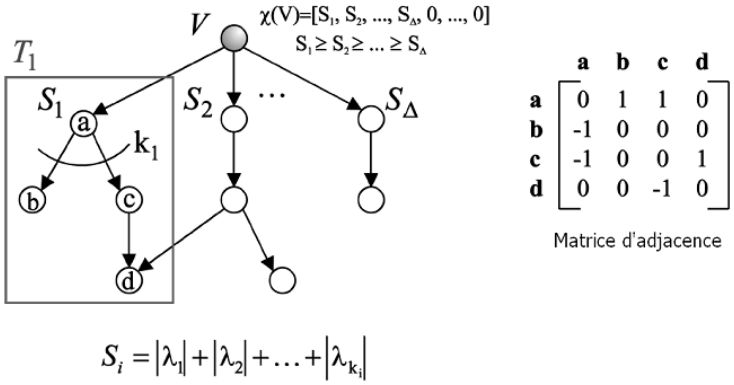


FIG. 5.4 – Calcul du vecteur de signature topologique d'un nœud [SMD+05].

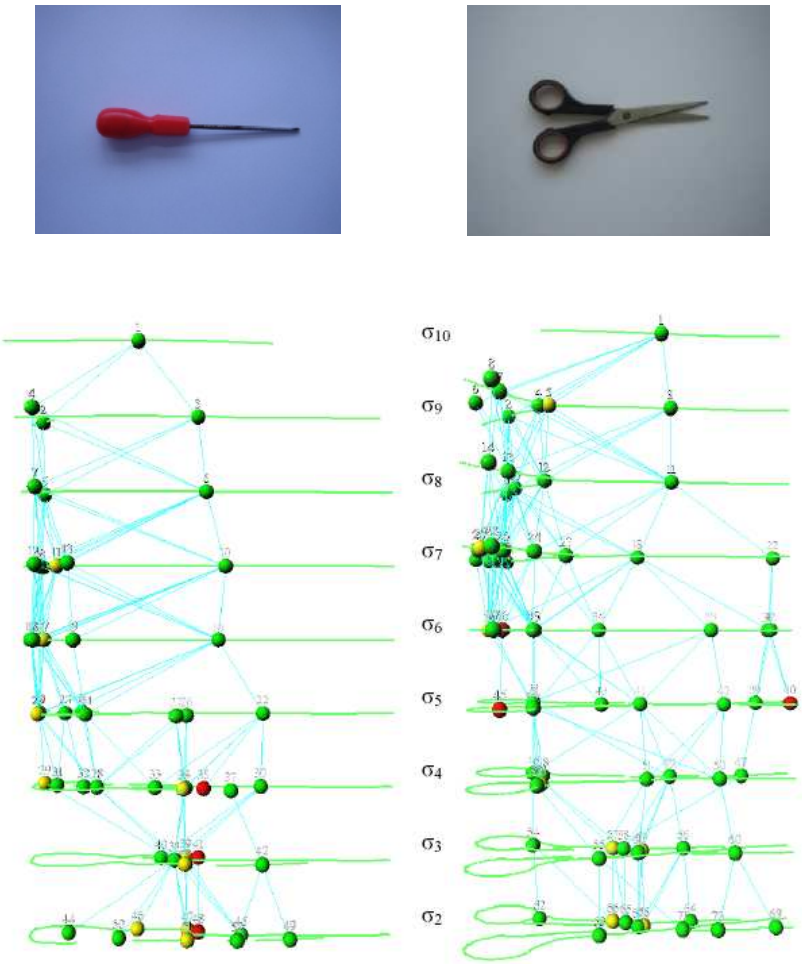


FIG. 5.5 – Objets tournevis, ciseau et les graphes de représentation correspondant.

L'évaluation de similarité de deux graphes se base sur la valeur totale de similarité par rapport au nombre de nœuds appariés. Nous détaillons la mesure de similarité des nœuds, la contrainte symbolique d'arc :

Contrainte symbolique : Chaque nœud correspond à une caractéristique crête, vallée, pic. Deux nœuds sont similaires seulement lorsque les caractéristiques correspondantes sont de même type. Note que cette condition est suffisante mais pas nécessaire.

Comme chaque nœud indique un type de caractéristique, le type d'arc est une paire R-R, R-V, R-T, V-V, V-B, V-T, B-B, B-T, T-T, etc où R, V, B, T signifient crête (R), vallée (V), pic (B), pic négatif (T). Deux arcs sont mis en correspondance si et seulement s'ils sont de même type.

Ces conditions sur le type de nœud ou d'arc sont des conditions symboliques. Elles sont exactes, permettant d'enlever rapidement des non-correspondances, et rendent donc la mise en correspondance plus rapide que la mise en correspondance des graphes où les nœuds sont appariés seulement sur la base de mesures numériques.

Dissimilarité des nœuds : La dissimilarité de deux nœuds est mesurée par la dissimilarité topologique et la dissimilarité géométrique. La dissimilarité topologique est mesurée par la distance euclidienne de deux vecteurs de signature topologique. La dissimilarité géométrique est mesurée par la différence de longueur, de l'échelle, de l'histogramme de direction.

5.2 Premiers résultats de construction de l'graphe et de reconnaissance

Nous avons réalisé les premières expérimentations de construction de graphe-mod'les, et de mise en correspondance de deux graphes. Les objets de test sont des objets simples comme un tournevis, des ciseaux, une gomme, une clé, une agrafe, un rasoir, etc. L'idée est de tester la robustesse de la représentation face au changement de la taille et de l'orientation. La figure 5.6 montre 8 images de 6 classes d'objets modèles et 28 images de 8 objets translatés ou tournés pour le test.

La figure 5.7 montre deux graphes construits à partir de crêtes détectées à 9 niveaux dans les images correspondantes. Chaque sphère représente le centre gravité de la crête ou pic (vert : crête, rouge : pic positif, jaune : pic négatif). Les liens sont désignés par les lignes bleus.

Nous pouvons constater, par exemple, que deux objets différents, le tournevis et les ciseaux, ont la même représentation à un niveau d'échelle suffisamment grand, où ils sont tous les deux représentés par une crête. La différence apparaît lorsqu'on descend dans la dimension d'échelle. Notons également que la représentation montre qu'un tournevis peut être considéré comme la moitié d'un ciseau.

L'algorithme de mise en correspondance hiérarchique présenté ci-dessus a été implémenté pour apparier deux graphes de représentation. Dans un premier temps, les objets test sont les objets d'apprentissage avec rotation et un léger changement d'échelle. Il y a peu de variété dans une classe d'objets. La figure 5.8 montre le processus d'apprentissage et d'appariement des graphes pour reconnaître des objets.

La figure 5.9 montre un exemple de mise en correspondance de deux tournevis qui sont identiques mais le deuxième a subi une rotation. Les sphères violettes montrent les nœuds mis en correspondance. Nous constatons que tous les nœuds caractéristiques sont appariés comme on s'y attend. La mise en correspondance est correcte non seulement pour les niveaux abstraits mais aussi pour les niveaux détaillés.

Le tableau 5.10 récapitule le résultat de mise en correspondance de 28 images avec 8 modèles appris. L'élément (i, j) du tableau représente la dissimilarité entre l'graphe modèle j et l'graphe de l'objet i. La reconnaissance attendue est que toutes les valeurs sur le diagonale du tableau soient les plus petites

possible par rapport aux valeurs dans une même ligne. Le tableau indique le résultat est satisfaisant. Presque tous les objets nouveaux sont reconnus correctement. Le taux de reconnaissance est de 93%. Ce test montre la stabilité de la représentation par structure de graphe par rapport au changement de luminosité, rotation et taille.

FIG. 5.6 – Première ligne : Images d'apprentissage. Deuxième ligne : Images de test.

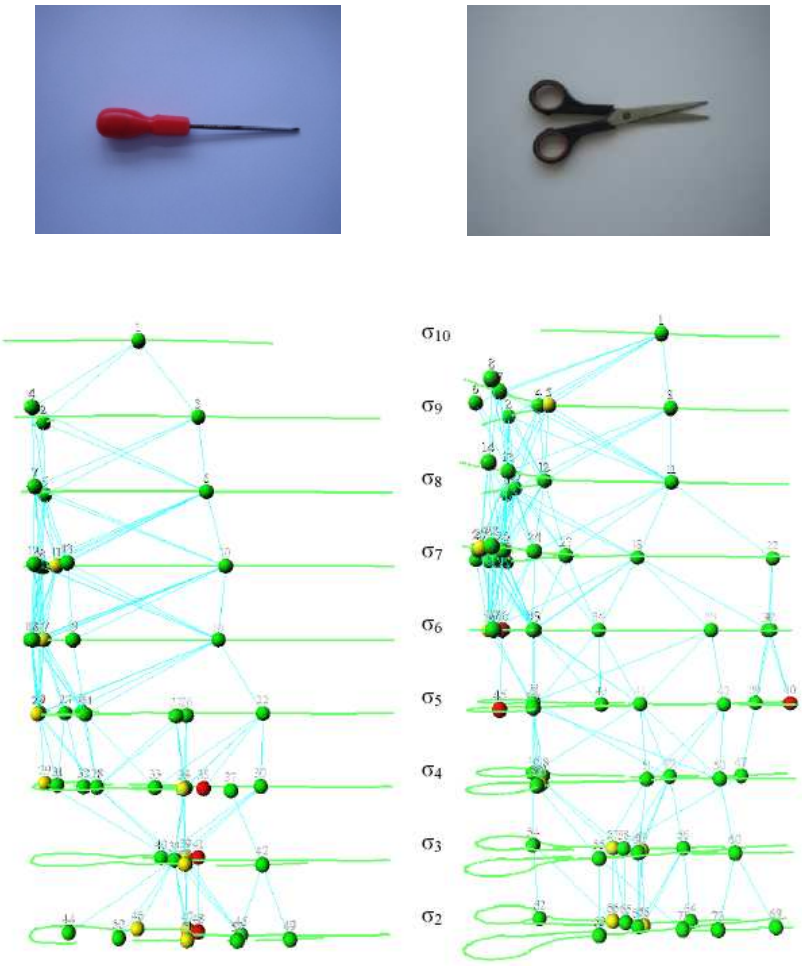


FIG. 5.7 – Objets tournevis, ciseau et les graphes de représentation correspondant

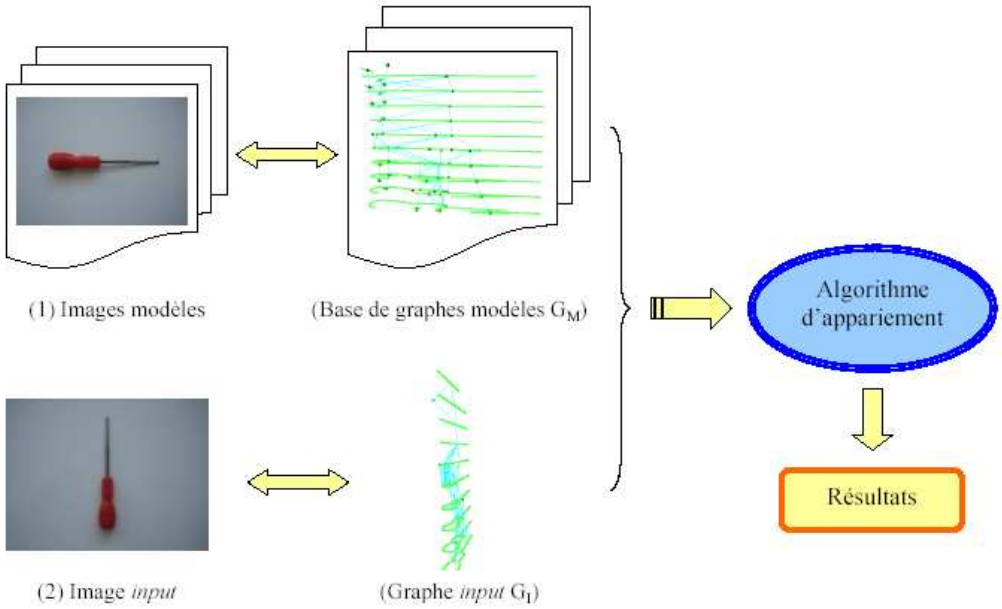


FIG. 5.8 – Mise en correspondance de deux graphes de représentation.

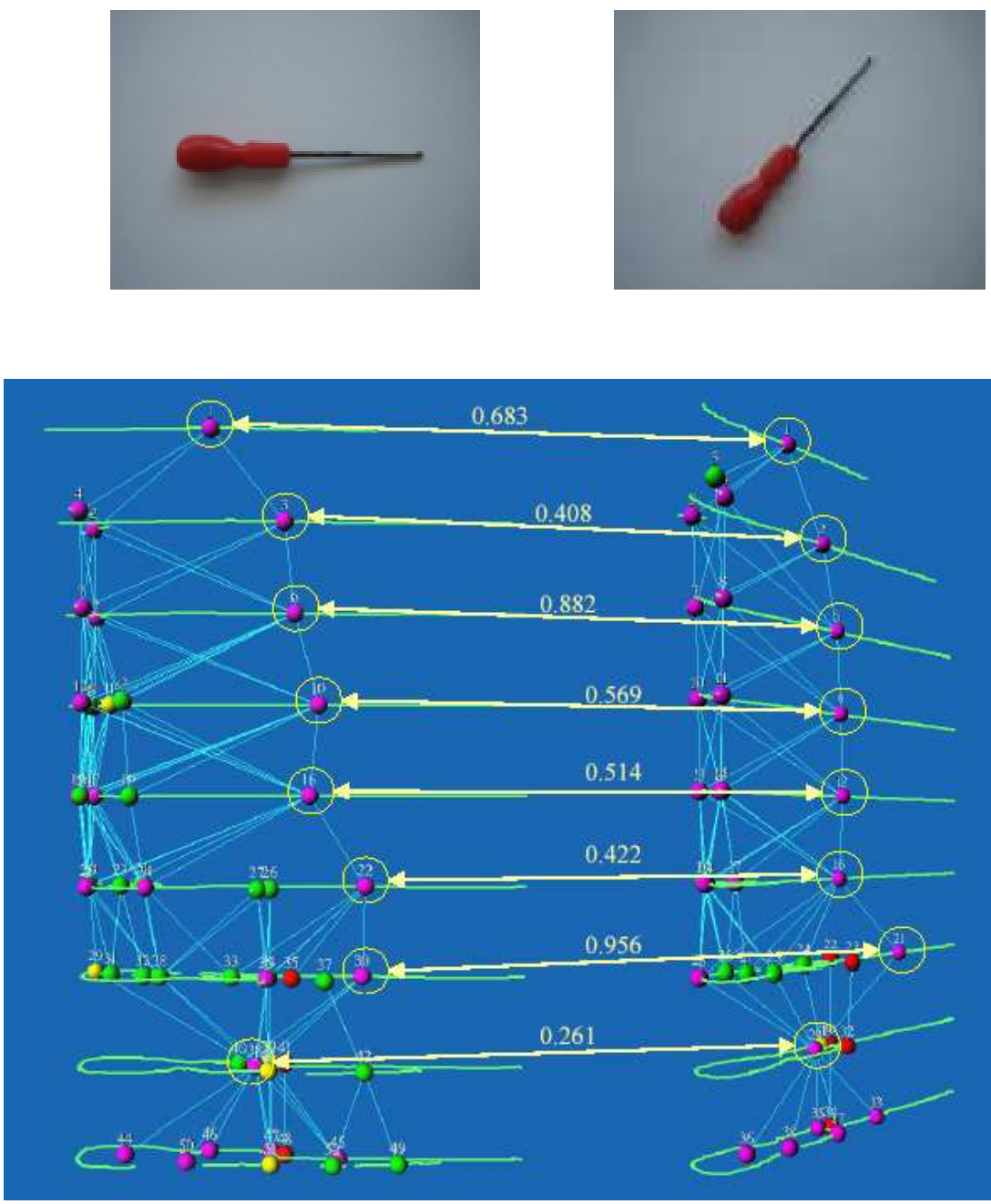


FIG. 5.9 – Mise en correspondance de deux tournevis.

I \ M	0	1	2	3	4	5	6	7	Min
0	0,552	2,691	1,277	17,308	1,52	2,95	1,694	2,948	0,552
1	0,901	3,54	1,28	14,177	1,255	8,093	2,623	2,817	0,901
2	0,635	1,88	0,943	13,269	1,838	6,788	1,398	2,586	0,635
3	1,888	1,192	34,981	2,885	2,042	29,551	39,093	44,117	1,192
4	10,668	0,906	61,802	5,493	2,292	2,875	69,968	78,001	0,906
5	64,464	0,633	2,989	1,289	2,52	2,122	66,779	77,36	0,633
6	8,998	0,548	3,671	1,908	2,12	2,983	59,457	70,815	0,548
7	76,49	0,792	4,76	1,604	68,102	2,824	58,65	71,904	0,792
8	8,073	7,127	1,544	7,747	1,779	4,422	2,308	5,31	1,544
9	2,641	4,943	1,031	5,935	2,975	3,572	1,685	3,492	1,031
10	7,652	7,965	2,941	38,464	3,992	32,707	4,11	33,161	2,941
11	25,351	32,856	2,595	37,96	10,245	34,126	4,487	27,687	2,595
12	115,166	3,341	6,307	2,214	103,486	3,436	92,27	116,837	2,214
13	92,129	1,886	71,194	0,524	84,016	2,51	6,294	96,597	0,524
14	11,019	8,384	4,029	3,91	6,682	1,779	2,305	39,045	1,779
15	12,943	32,663	13,372	39,381	3,624	34,919	22,249	29,211	3,624
16	1,08	18,925	1,697	23,988	0,674	20,353	2,142	6,479	0,674
17	8,805	1,631	1,252	24,841	1,161	2,809	1,792	3,629	1,161
18	81,546	2,331	6,559	1,916	73,316	1,065	4,355	87,085	1,065
19	89,787	2,884	13,071	2,323	81,778	0,546	5,237	98,765	0,546
20	120,47	26,751	96,885	1,534	112,361	1,498	91,832	126,707	1,498
21	4,25	40,702	2,321	11,533	4,019	3,81	0,865	4,128	0,865
22	5,208	4,752	1,918	6,483	2,778	9,72	0,863	4,349	0,863
23	3,836	3,245	1,165	6,696	2,203	4,881	0,396	4,479	0,396
24	3,332	3,007	1,261	8,425	3,229	3,968	0,742	6,191	0,742
25	3,979	26,98	2,641	31,573	2,298	27,597	1,627	1,56	1,56
26	3,682	19,996	1,917	20,651	4,558	18,232	2,569	1,583	1,583
27	5,978	23,658	4,045	25,995	5,66	23,963	3,052	0,581	0,581

tournevis n°1 tournevis n°2 ciseau gomme agrafe clé n°1 clé n°2 rasoir

FIG. 5.10 – Résultat de mise en correspondance des objets de test avec des objets d'apprentissage.

Chapitre 6

Conclusions et Perspectives

6.1 Conclusions

Pour répondre à la question *quels rôles peuvent jouer les lignes d'intérêt naturelles pour la représentation d'objets*, nous avons mené les recherches sur la définition mathématique de lignes d'intérêt naturelles, sa détection et ses applications à la modélisation des objets.

6.1.1 Une exploration vers une nouvelle caractéristique

Malgré l'abondance des caractéristiques existantes dans la littérature, le fait d'introduire un nouvelle caractéristique - la ligne d'intérêt naturelle - n'est pas inutile. Permettant de représenter génériquement toutes les structures allongées ou arrondies, elle enrichit le jeu de caractéristiques, afin de pouvoir modéliser toutes les informations présentes dans une image.

En introduisant les lignes d'intérêt naturelles, nous avons proposé une nouvelle méthode de détection des points de crêtes basée sur le Laplacien de Gaussien. Cette méthode a deux avantages : elle localise plus correctement les points de crête et offre la possiblité de mesurer la taille des structures, ce qui n'a jamais été évoqué par les méthodes existantes.

En outre, nous avons étudié les crêtes dans l'espace-échelle. Cette étude nous permet de détecter des lignes de crête de tailles différentes, correspondant à des structures de tailles différentes dans une image. Elle permet en particulier une représentation grossière-détaillée d'un même objet. Celle-ci est vraiment utile pour une reconnaissance générique, où on s'intéresse premièrement à une caractérisation abstraite de l'objet.

Notre méthode de détection de crête a expérimentalement montré être plus robuste au bruit que les méthodes basées directement sur le signal ou sur sa courbure. Elle a produit des crêtes plus continues, ce qui est une propriété importante pour des caractéristiques de type ligne.

6.1.2 Une toute nouvelle méthode de détection de texte

La première application de l'utilisation des crêtes, la détection de texte, a obtenu un résultat parfaitement surprenant. Nous nous sommes basés sur le changement de structure d'un texte à deux niveaux d'échelles pour modéliser une ligne de texte et la distinguer d'autres types d'objets.

Un point tout nouveau dans cette approche est la considération d'une ligne de texte en tant qu'objet structurel simple, dont les crêtes à deux niveaux sont suffisantes pour sa description. Cette modélisation a été expérimentée avec plusieurs types de texte. Sa performance s'est montrée comparable avec celle des méthodes de l'état de l'art. Nous sommes très confiants en ce que la détection de texte obtiendra une meilleure performance encore, si les crêtes sont détectées à une bonne échelle.

6.1.3 Une représentation de personne à base de crête

Les textes sont des "objets artificiels", composés de structures allongées qui sont parfaitement représentées par les lignes de crête. Les objets dans le monde réel sont plus complexes qu'un texte, donc plus difficile à modéliser. Il n'est donc pas évident d'y reproduire le même succès. Le deuxième cas d'étude du rôle des crêtes a été mené sur la détection de personnes, une application très attractive mais aussi très polémique à l'heure actuelle.

Nous avons modélisé la forme d'une personne par des crêtes principales représentant ses membres : le torse et les jambes. La configuration de ces membres indique dans quel état est la personne (en mouvement ou immobile). Malgré la simplicité de cette modélisation, elle est suffisamment efficace pour détecter des personnes dans une scène où la confusion avec d'autres objets peut facilement poser problème. En outre, elle est applicable à l'évaluation du nombre de personnes dans un groupe.

6.2 Perspectives

La thèse a atteint son première objectif : " montrer les rôles d'une crête pour la représentation de structures dans l'image via des exemples concrets". Nous avons étudié la détection de crêtes sur des objets simples - les textes (chapitre 4), puis sur des objets plus complexes - les personnes (chapitre 3), pour finalement proposer une ébauche de généralisation à tout type d'objet quotidien (chapitre 5).

Nous ne voulons pas nous arrêter là. Avec toutes les belles propriétes de crêtes étudiées dans l'espace-échelle, nous sommes à même de croire qu'une représentation à base de crêtes sera précieuse pour la classification d'objets.

Le test de reconnaissance d'objets simples basée sur le graphe attribué relationnel a montré la stabilité de la méthode de représentation et l'efficacité de la mise en correspondance hiérarchique. Pour construire un système de reconnaissance d'objets générique, nous prévoyons plusieurs améliorations possibles :

- *Construction d'un vocabulaire informatif et discriminant :* Les crête et les pics détectés à plusieurs échelles constituent un vocabulaire. Comme certaines crêtes et pics, représentant la même structure, peuvent se repéter à plusieurs échelles, ceci rend le vocabulaire redondant et la mise en correspondance ambiguë. Une solution pour réduire la redondance est de décrire les crêtes ou les pics par des mesures numériques (ie. vecteur de descripteurs) multimodal[1] puis de les classer en petit nombre de classes. Le vocabulaire est alors composé des noyaux de chaque classe, qui sont plus discriminants que l'ensemble de crêtes et pics d'origine. Ce vocabulaire peut compléter les vocabulaires existants construits avec des points d'intérêt [DS05, LDJ06]. En construisant le vocabulaire de cette manière, deux problèmes sont résolus. D'abord, comme les fausses crêtes ne sont pas des caractéristiques intrinsèques des objets appris, elles ne sont pas stables. L'élimination des fausses crêtes peut s'effectuer en ne considérant pas ces classes. Ensuite, le regroupement des

[1] un vecteur de descripteurs de différents aspects : shape context, histogramme de couleur ou de champs réceptifs, etc.

caractéristiques similaires dans une classe et l'utilisation du noyau de la classe comme représentant permet de ne pas avoir à sélectionner l'échelle caractéristique d'une crête. La détection de celle-ci est en effet un problème très délicat parce qu'elle peut varier fortement le long d'une ligne de crête.

- *Apprentissage incrémental du modèle :* Le graphe de représentation présenté ci-dessus est construit pour chaque objet. Pour la reconnaissance de classes d'objets, plusieurs graphes représentant des objets d'une même classe doivent être appris et stockés. La reconnaissance consiste alors à comparer le modèle de l'objet avec MxN modèles dans la base, où M est le nombre d'objets exemples dans chaque classe et N le nombre de classes. Ce problème est MxN fois plus coûteux que le problème d'appariement de deux graphes. Une idée est de construire une représentation unique pour chaque classe d'objets. L'algorithme de construction est donc incrémental et capable d'insérer des nouvelles caractéristiques avec des nouvelles relations parmi celles existantes. Ainsi, la mise en correspondance peut être un processus d'insertion de la nouvelle structure parmi les structures modèles.
- *Modélisation de personnes de façon plus sophistiquée :* Nous souhaitons appliquer la représentation proposée pour la modélisation de personnes de façon plus sophistiquée. Avec la représentation par un graphe (voir figure 6.1), nous espérons pouvoir augmenter la séparation entre la classe de personnes et celle de non-personnes. L'apprentissage de vocabulaire permet de déterminer facilement les crêtes torse, jambes, etc. sans avoir besoin de l'orientation de la personne, comme c'est le cas pour l'approche présentée dans le chapitre 3.

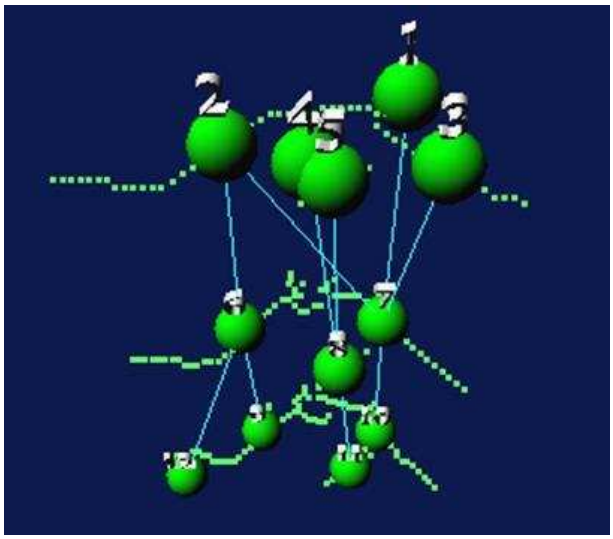


FIG. 6.1 – (a) Image d'une personne. (b) Représentation de personne par structure d'graphe.

Bibliographie

[AC99] J. K. Aggrawal and Q. Cai. Human motion analysis : A review. *Computer Vision and Image Understanding*, 73(3) :428–440, 1999.

[aQH97] Y. T. Cui anf Q. Huang. Character extraction of license plates from video. In *Proc. of International Conference on Computer Vision and Pattern Recognition*, pages 502–, 1997.

[AR02] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proc. of ECCV02*, 2002.

[Bau95] A. M. Baumberg. *Learning Deformable Models for Tracking Hu man Motion*. PhD thesis, University of Leeds, School of Computer Studies, October 1995.

[BBD+02] A. Bishnu, P. Bhowmick, J. Dey, B. B. Bhattacharya, M. K. Kundu, C. A. Murthy, and T. Acharya. Combinatorial classification of pixels for ridge extraction in a gray-scale fingerprint image. In *Proc. 3rd India Conference on Computer Vision Graphic and Image Processing (ICVGIP'02)*, pages 451–456, Ahmedabad India, December 2002.

[BD01] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(3) :257–267, March 2001.

[BE91] A. Bengtsson and J. Eklundh. Shape representation by multiscale contour approximation. *IEEE Transactions on PAMI*, 13 :85–93, 1991.

[BGT96] J. W. Bruce, P. J. Giblin, and F. Tari. Ridges, crests and sub-parabolic lines of evolving surfaces. *International Journal of Computer Vision*, 18(3) :195–210, 1996.

[BH93] A. M. Baumberg and D. C. Hogg. Learning flexible models from image sequences. Technical report, University of Leeds, October 1993.

[BH94] A. M. Baumberg and D. C. Hogg. An efficient method for contour tracking using active shape models. Technical report, University of Leeds, April 1994.

[Bie85] I. Biederman. Human image understanding : Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32 :29–73, 1985.

[BL93] R. Bergevin and M. D. Levine. Generic object recognition : Building and matching coarse descriptions from line drawings. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(1) :19–36, Jan 1993.

[Blu67] H. Blum. A transformation for extracting new descriptors of shape. *In W. Wathen-Dunn(editor), Models for the Perception of Speech and Visual Form, MIT Press, Cambridge Mass.*, 1967.

[BM00] S. Belongie and J. Malik. Matching with shape contexts. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, June 2000.

[BMP02] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape context. *IEEE Transactions on Pattern Analysis and Machine Learning*, 24(24) :509–522, 2002.

[BPK98] A. G. Belyaev, A. A. Pasko, and T. L. Kunii. Ridges and ravines on implicit surfaces. In *Proc. of Computer Graphics International '98*, pages 530–535, Hannover, June 1998.

[Bro83] R. Brooks. Model-based 3-d interpretations of 2-d images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 5(2) :140–150, 1983.

[BWBD86] J. Babaurd, A. P. Witkin, M. Baudin, and R. O. Duda. Uniqueness of the gaussien filter for scale-space filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1) :26–33, 1986.

[BWP98] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. ECCV*, pages 628–641, 1998.

[BY92] A. Blacke and A. Yuille. Deformable contours : Modeling, extraction. *Active Vision, MIT Press, Cambridge, MA*, 1992.

[Can83] J. Canny. A computational approach to edge detection. Master's thesis, MIT, 1983.

[CCBS97] J. Coutaz, J. L. Crowley, F. Berard, and D. Salber. Eigenspace coding as a means to support privacy in computer mediated communication. *Interact*, 1997.

[CdVC99] O. Chomat, V. Colin de Verdiere, and J. L. Crowley. Recognizing goldfish ? or local scale selection for recognition technique. In *International Symposium for Intelligent Robotics System*, pages 197–206, 1999.

[Che03] D. Chen. *Text Detection and Recognition in Images and Video Sequences*. PhD thesis, Institut Dalle Molle d'Intelligence Artificielle Perceptive(IDIAP), 2003.

[CHRC04] A. Caporossi, D. Hall, P. Reignier, and James L. Crowley. Robust visual tracking from dynamic control of processing. In *Performance and Evaluation of Tracking and Surveillance PETS'04*, 2004.

[CLK+00] Robert Collins, Alan Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, and Osamu Hasegawa. A system for video surveillance and monitoring. Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2000.

[CM00a] P. Clark and M. Mirmehdi. Location and recovery of text on oriented surfaces. In *SPIE Conference on Document Recognition and Retrieval VII*, pages 267–277, January 2000.

[CM00b] Paul Clark and Majid Mirmehdi. Finding text regions using localised measures. In Majid Mirmehdi and Barry Thomas, editors, *Proc. of the 11th British Machine Vision Conference*, pages 675–684. BMVA Press, September 2000.

[CMPC06] A.I. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless tracking for augmented reality : the virtual visual servoing framework. *IEEE Trans. on Visualization and Computer Graphics*, 12(4) :615–628, July 2006.

[CP80] P. Chen and T. Pavlidis. Image segmentation as an estimation problem. *Computer Graphics and Image Processing*, 12 :153–172, 1980.

[CP84] James L. Crowley and Alice C. Parker. A representation for shape based on peaks and ridges in the difference of low-pass transform. *IEEE Transactions on Pattern analysis and Machine Intelligence*, 6(2) :145–169, March 1984.

[CR03] J. L. Crowley and O. Riff. Fast computation of characteristic scale using a half octave pyramid. In *Scale-space 03, 4th International Conference on Scale-Space theories in Computer Vision*, Isle of Skye, Scotland, UK, June 2003.

[Cro81] J. L. Crowley. *A representation for Visual Information*. PhD thesis, Carnegie Mellon University, the Robotics Institute, 1981.

[CTCS94] P. Chung, C. Tsai, E. Chen, and Y. Sun. Polygonal approximation using a competitive hopfield neural network. *Pattern Recognition*, 27 :1505–1512, 1994.

[CTS95] S. Chinveeraphan, R. Takamatsu, and M. Sato. Understanding of ridge-valley lines on image-intensity surfaces in scale-space. In *Proc. of International Conference CAIP'95, Computer Analysis of Images and Patterns*, Prague, Czech Republic, September 1995.

[DS05] G. Dorko and C. Schmid. Object class recognition using discriminative local features. Technical Report 5497, Institut National de Recherche en Informatique et Automatique, 2005.

[dV99] V. Colin de Verdiere. *Representation et Reconnaissance d'Objet par Champs Receptifs*. PhD thesis, INPG, 1999.

[Ebe94] D. Eberly. *Geometric Methods for Analysis of Ridges in N-Dimensional Images*. PhD thesis, North Carolina University, USA, 1994.

[Ebe96] D. Eberly. *Ridges in Image and Data Analysis*. Kluwer Academic Publishers, 1996.

[EF86] M. A. Eshera and K. S. Fu. An image understanding system using attributed symbolic representation and inexact graph-matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(5) :604–617, 1986.

[EGM+94] D. Eberly, R. Gardner, B. Morse, S. Pizer, and C. Scharlach. Ridges for image analysis. *Journal of Mathematical Imaging and Vision*, 4 :353–373, 1994.

[FA91] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 13(9) :891–906, 1991.

[FF95] B. V. Funt and G. D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 17(5) :522–529, 1995.

[FKVL99] A. Farina, Zs. M. Kovács-Vajna, and A. Leone. Fingerprint minutiae extraction from skeletonized binary. *Pattern Recognition*, 32 :877–889, 1999.

[FL98] H. Fujiyoshi and Alan J. Lipton. Real-time human motion analysis by image skeletonization. In *Proc. of the Workshop on Application of Computer Vision*, October 1998.

[FM81] K. Fu and J. Mui. A survey on image segmentation. *Pattern Recognition*, 13 :3–16, 1981.

[FPZ03] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2003.

[FPZ04] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV04*, 2004.

[Fre61] H. Freeman. On the encoding of arbitrary geometric configuration. *IRE Transactions*, 1961.

[FtHRKV92] L. M. J. Florack, B. M. ter Haar Romeney, J. J. Koenderink, and M. A. Vierger. Scale and the differential structure of images. *Image and Vision Computing*, 10 :376–388, July/August 1992.

[Gil98] S. Gilles. *Robust description and matching of images*. PhD thesis, University of Oxford, 1998.

[Gos85] A. Goshtasby. Description and discrimination of planar shapes using shape matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(15) :635–646, June 1985.

[GP93] John M. Gauch and Stephen M. Pizer. The intensity axis of symmetry and its application to image segmentation. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 15(8) :753–770, August 1993.

[GR96] S. Gold and A. Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(4) :377–388, 1996.

[Hal01] D. Hall. *Viewpoint independent Recognition of Objects from Local Appearance*. PhD thesis, INPG, 2001.

[Hal04] D. Hall. A system for object class detection. *Cognitive Vision System*, 2004.

[HAMMC05] H. Hadj Abdelkader, Y. Mezouar, P. Martinet, and F. Chaumette. Asservissement visuel en vision omnidirectionnelle à partir de droites. *Traitement du Signal*, 22(5) :462–482, September 2005.

[Har83] Robert M. Haralick. Ridges and valleys on digital images. *Computer Vision, Graphics, And Image Processing*, 22 :28–38, 1983.

[Har99] *Hydra : Multiple People Detection and Tracking Using Silhouettes*, Fort Collins, Colorado, 26 June 1999.

[HCCdV00] D. Hall, J. L. Crowley, O. Chomat, and V. Colin de Verdière. View invariant object recognition using coloured receptive fields. *Machine Graphics and Vision*, 9(2) :341–352, June 2000.

[HdVC00] D. Hall, V. Colin de Verdière, and J. L. Crowley. Object recognition using coloured receptive field. In *Proc. of 6th European Conference on Computer Vision*, pages 164–178, Springer Verlag, Dublin, June 2000.

[HHD98a] I. Haritaoglu, D. Harwood, and L. S. Davis. Ghost : A human body part labeling system using silhouettes. Brisbane, Australia, August 16-20 1998.

[HHD98b] I. Haritaoglu, D. Harwood, and L. S. Davis. W4 :who ? when ? where ? what ? a real time system for detecting and tracking people. In *Proc. ofInternational Conference on Face and Gesture Recognition*, April 14-16 1998.

[HHD99] I. Haritaoglu, D. Harwood, and L. S. Davis. Hydra : Multiple people detection and tracking using silhouettes. In *Proc. of 10th International Conference on Image Analysis and Processing (ICIAP'99)*, pages 280–, 1999.

[HK00] Y. M. Y. Hasan and L. J. Karam. Morphological text extraction from images. *IEEE Transactions on Image Processing*, 9(11) :1978–1983, 2000.

[HLS02] D. Hall, B. Leibe, and B. Schiele. Saliency of interest points under scale changes. In *Proc. of British Machine Vision Conference*, September 2002.

[HMB78] S. Hsu, J. L. Mundy, and P. R. Beaudet. Web representation of image data. In *Proceedings of International Joint Conference on Pattern Recognition*, pages 675–680, Kyoto, Japan, 1978.

[Hol92] J. Hollingum. Automated fingerprint analysis offers fast verification. *Sensor Review*, 12(13) :12–15, 1992.

[HPRC04] D. Hall, F. Pelisson, O. Riff, and J. L. Crowley. Brand identification using gaussian derivative histograms. *Machine Vision and Applications*, 2004.

[HS85] R. Haralick and L. Shapiro. Image segmentation techniques. *Computer Vision, Graphics, Image Processing*, 1985.

[HS94] G. Healey and D. Slater. Using illumination invariant color histogram descriptors for recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 355–360, 1994.

[Hu62] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8 :179–187, 1962.

[HU90] D. Huttenlocher and S. Ullman. Recognizing solid object by alignment with an image. *International Journal of Computer Vision*, 5(2) :195–212, 1990.

[Hun94] M. Hunke. Locating and tracking of human faces with neural networks. Technical report, Carnegie Mellon University, 1994.

[IM82] Y. Ikebe and S. Miyamoto. Shape design, representation, and restoration with splines. *Picture Engineering*, pages 75–95, 1982.

[JB92] A. K. Jain and S. Bhattacharjee. Text segmentation using gabor filters for automatic document processing. *Machine Vision and Application*, 5 :169–184, 1992.

[JBAM96] M. A. Viergever J. B. A. Maintz, P. A. van den Elsen. Evaluation of ridge seeking operators for multimodality medical image matching. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, pages 353–365, April 1996.

[JHPB97] A. K. Jain, L. Hong, S. Pankanti, and R. Bolle. An identity-authentification system using fingerprints. In *Proc. of IEE*, volume 85, pages 1365–1388, 1997.

[JJKK99] K. Y. Jeong, J. Jung, E. Y. Kim, and H. J. Kim. Neural network-based text location for news video indexing. In *Proc. of IEEE International Conference on Image Processing*, volume 3, pages 319–323, 1999.

[JK96] A. K. Jain and K. Karu. Learning texture discrimination masks. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(2) :195–205, February 1996.

[Jun01] K. Jung. Neural network-based text location in color images. *Pattern Recognition Letters*, 22(14) :1503–1515, 2001.

[kB01] T. kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, pages 83–105, 2001.

[KD05] Y. Keselman and S. Dickinson. Generic model abstraction from examples. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(7) :1141–1156, 2005.

[KJ04] A. K. Jain K. Jung, K. I. Kim. Text information extraction in images and videos : A survey. *Pattern Recognition*, 37(5) :977–997, May 2004.

[KStHRV01] Stiliyan N. Kalitzin, Joes Staal, Bart M. ter Haar Romeny, and Max A. Viergever. A computational method for segmenting topological point-sets and application to image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(5) :447–459, 2001.

[KV99] R. Khardon and L. G. Valiant. Relational learning for nlp using linear threshold elemen. In *Proc. of IJCAI'99*, pages 911–919, 1999.

[KvD84] J. J. Koenderink and A. J. van Doorn. The structures of images. *Biological Cybernetics*, (50) :363–370, 1984.

[KvD87] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, (55) :367–375, 1987.

[Lai94] K. F. Lai. *Deformable Contours : Modeling, Extraction, Detection and Classification*. PhD thesis, Electrical and Computer Engineering Department, University of Wisconsin at Madison, 1994.

[LBBK97] V. Lang, A. G. Belyaev, I. A. Bogaevsici, and T.L. Kunii. Fast algorithms for ridge detection. In *Proceedings of International Conference on Shape Modeling and Applications (SMA'97)*, pages 189–197, Aizu-Wakamatsu, Japan, March 1997.

[LDJ06] D. Larlus, G. Dorko, and F. Jurie. Création de vocabulaires visuelles efficaces pour la catégorisation d'images. In *Congrès de Reconnaissance des formes et Intéligence Artificielle*, 2006.

[LFP98] A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target classification and tracking from real-time video. In *DARPA Image Understanding Workshop (IUW'98)*, 1998.

[Lie96] R. Lienhart. Indexing and retrieval of digital video sequences based on automatic text recognition. In *Fourth ACM International Multimedia Conference*, Boston, USA, November 1996.

[Lin94] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[LL00] I. Laptev and T. Lindeberg. Tracking of multi-state hand models using particle filtering and a hierarchy of multi- scale image features. Technical report, Computational Vision and Active Perception Laboratory (CVAP), 2000.

[LLSV99] Antonio M. López, F. Lumbreras, J. Serrat, and J. J. Villaneuva. Evaluation of methods for ridge and valley detection. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 21(4) :327–335, April 1999.

[Lon98] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8) :983–1001, 1998.

[Low85] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic, 1985.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004.

[LS96] R. Lienhart and F. Stuber. Automatic text recognition in digital videos. In *Image and Video Processing IV 1996, SPIE 2666-20*, 1996.

[LS03] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, USA, June 2003.

[LW02] R. Lienhart and A. Wernicke. Localizing and segmenting text in images and videos. *IEEE Transactions on Circuits and System for Video Technology*, 12(4) :256–268, April 2002.

[LY95] M. K. Leung and Y. H. Yang. First sight : A human body outline labeling system. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(4) :359–377, April 1995.

[MA88] A. J. Maren and M. Ali. Hierarchical scene structure representations to facilitate image understanding. *ACM*, 1988.

[MAM95] O. Monga, N. Armande, and P. Montesinos. Thin nets and crest lines : Application to satellite data and medical images. Technical Report 2480, INRIA, February 1995.

[Mar82] D. Marr. *Vision : A computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company Sans Francisco, 1982.

[MBF92] O. Monga, S. Benayoun, and D. Faugeras. Using partial derivatives of 3d images to extract typical surface features. In *Proc. of IEEE Conference on Vision and Pattern Recognition*, Urbana Champaign, Illinois, July 1992.

[MBM01] G. Mori, S. Belongie, and J. Malik. Shape context enable retrieval of similar shapes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[MC95] J. Martin and J. L. Crowley. Comparison of correlation technique. In *Intelligent Autonomous Systems,IAS'95*, pages 86–93, Karlsruhe, Germany, 1995.

[MCLS02] W. Mao, F. Chung, K. Lanm, and W. Siu. Hybrid chinese/english text detection in images and video frames. In *Proc. of International Conference on Pattern Recognition*, volume 3, pages 1015–1018, 2002.

[MF99] J. Miller and J. Furst. The maximal scale ridge incorporating scale into the ridge definition. In *Lecture Notes in Computer Science*, volume 1682, pages 93–105, 1999.

[MHSTS05] H. Mength, D. R. Hardoon, J. Shawe-Taylor, and S. Szedmak. Generic object recognition by combining distinct features in machine learning. pages 90–98, 2005.

[MM99] S. Messelodi and C. M. Modena. Automatic identification and skew estimation of text lines in real scene images. *Pattern Recognition*, 32(5) :789–808, 1999.

[MN95] H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal on Computer Vision*, 1995.

[MPL93] B. Morse, S. Pizer, and A. Liu. Multiscale medial analysis of medical imaging. In *Proc. of International Conference on Information Processing in Medical Imaging*, volume 687, pages 112–131, Flagstaff, Arizona, USA, 1993.

[MS01] K. Mikolajczuk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV01*, pages 525–531, July 2001.

[MS04a] K. Mikolajczyk and C. Schmid. Comparison of affine-invariant local detectors and descriptors. In *Proc. of 12th European Signal Processing Conference*, 2004.

[MS04b] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1) :63–86, 2004.

[MS05] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10) :1615–1630, 2005.

[NTG+05] A. Negre, H. Tran, N. Gourier, D. Hall, A. Lux, and J. L. Crowley. Object recognition invariant to viewpoint. Technical report, Deliverable CAVIAR, 2005.

[OBS04] Y. Ohtake, A. Belyaev, and H. P. Seidel. Ridge-valley lines on meshes via implicit surface fitting. In *Proc. of the 2004 SIGGRAPH Conference*, volume 23, pages 609–612, August 2004.

[OFPA04a] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Generic object recognition with boosting. Technical report, Graz University of Technology, 2004.

[OFPA04b] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV'04*, pages 71–84, 2004.

[OI96] K. Ohba and K. Ikeuchi. Recognition of the multi specularity objects using the eigen window. In *Proc. International Conference on Pattern Recognition*, 1996.

[OI97] K. Ohba and K. Ikeuchi. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9), 1997.

[OSA94] J. Ohya, A. Shio, and S. Akamatsu. Recognizing characters in scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2) :214–220, February 1994.

[Pat75] K. Paton. Picture description using legendre polynomials. *Computer Graphics and Image Processing*, (4) :40–54, 1975.

[PC02a] J. Piater and J. L. Crowley. Event-based activity analysis in live video using a generic object tracker. pages 1–8, 2002.

[PC02b] J. Piater and J. L. Crowley. Multi-modal tracking of interacting targets using gaussian approximations. 2002.

[PCpS99] J. Ponce, M. Cepeda, S. pae, and S. Sullivan. Shape models and object recognition. In *Shape, Contour Grouping in Computer Vision, Springer*, 1999.

[Pen86] A. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28 :293–331, 1986.

[Pen99] Antonio M. López Pena. *Multilocal Methods for Ridge and Valley Delineation in Image Analysis*. PhD thesis, Universita Autònoma de Barcelona, August 1999.

[Pha05] T. Pham. Methode de mise en correspondance hierachique en reconnaissance d'objet. Master's thesis, Insitut National Polytechnique de Grenoble, 2005.

[PHRC03] F. Pelisson, D. Hall, O. Riff, and J. Crowley. Brand identification using gaussian derivative histograms. In *Proc. of International Conference on Vision Systems*, Graz, Austria, April 2003.

[PM83] S. Parui and D. Majumder. Symmetry analysis by computer. *Pattern Recognition*, 16 :63–67, 1983.

[PM04] M. Pressigout and E. Marchand. Model-free augmented reality by virtual visual servoing. In *IAPR Int. Conf. on Pattern Recognition, ICPR'04*, volume 2, pages 887–891, Cambridge, UK, August 2004.

[Pou98] F. Pourraz. Estimation de position d'un robot mobile par projection dans un espace de composantes principales. Master's thesis, ENSIMAG, 1998.

[PP93] N. Pal and S. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26 :1277–1294, 1993.

[PPP98] L. Paletzta, M. Prantl, and A. Pinz. Reinforcement learning for autonomous three-dimensional object recognition. In *Symposium on Intelligent Robotics Systems, SIRS 98*, pages 63–81, Edinburgh, United Kingdom, 1998.

[PR92] R. J. Prokop and A. P. Reeves. A survey of moment-based techniques for unoccluded object representation and recognition. In *Proc. of CVGIP : Graphical Models and Image Processing*, pages 438–460, 1992.

[PSL95] E. Piegay, N. Selmaoui, and C. Leschi. Crest line detection by valleys spreading. In Václav Hlavác Radim Sára, editor, *Proc. of International Conference on Pattern Recognition*, 1995.

[PSZ99] M. Pelillo, K. Siddiqi, and S. W. Zucker. Matching hierarchical structures using association graphs. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(11) :1105–, November 1999.

[RA01] P. S. Rodrigues and A. A. Araujo. A region-based object recognition algorithm. 2001.

[RB95] R. P. N. Rao and D. H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence Journal*, (78) :461–505, 1995.

[RFZ05] D. Ramanan, D. Forsyth, and A. Zisserman. Tracking people and recognizing their activities. In *Video Proceedings of Computer Vision and Pattern Recognition*, San Diego, CA, June 2005.

[RLSP05] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision.*, 2005.

[RM93] H. Rom and G. Medioni. Hierarchical decomposition and axial shape description. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(10) :973–981, 1993.

[Roo] Root image processing lab. http ://rootimage.msu.edu.

[SAN+04] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. V. Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4) :501–509, April 2004.

[SBK99] K. Sobottka, H. Bunke, and H. Kronenberg. Identification of text on colored book and journal covers. In *Proc. of International Conference on Document Analysis and Recognition*, pages 57–62, Bangalore, India, September 1999.

[SC96a] B. Schiel and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *Proc. of ECCV'96*, 1996.

[SC96b] B. Schiel and J. L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *Proc. of ICPR'96*, 1996.

[SC96c] B. Schiele and James L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV-'96, European Conference on Computer Vision*, Cambridge UK, apr 1996.

[SC00] B. Schiele and J. L. Crowley. Object recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1) :31–50, 2000.

[Sch92] C. Schmid. *Appariement d'image par invariants locaux de niveaux de gris*. PhD thesis, Institut National Polytechnique de Grenoble, 1992.

[SK87] I. Sirovich and M. Kirby. Low-dimensional procedure for the characterisation of human faces. *J. Opt. Soc. Am*, 4(3) :519–524, 1987.

[SK95] M. Smith and T. Kanade. Video skimming for quick browsing based on audio and image characterization. Technical Report CMU-CS-95-186, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, July 1995.

[SKA+02] J. Staal, S. Kalitzin, M. D. Abrámoff, T. Berendschot, B. van Ginneken, and M. A. Viergever. Classifying convex sets for vessel detection in retinal images. In *Proc. of International Symposium on Biomedical Imaging*, pages 269–272, 2002.

[SKC02] B. Sin, S. Kim, and B. Cho. Locating characters in scene images using frequency features. In *Proc. of International Conference on Pattern Recognition*, volume 3, pages 489–492, 2002.

[SKK01] T. B. Sebatian, P. N. Klein, and B. B. Kimia. Recognition of shapes by editing shock graphs. In *Proc. of International Conference on Computer Vision*, pages 755–762, 2001.

[SLE93] N. Selmaoui, C. Leschi, and H. Emptoz. Crest line detection in grey level images : Studies of different approaches and proposition of a new one. In Dimitry Chetverikoy and Walter G. Kropatsh, editors, *Proc. of International Conference in Computer Analysis of Images and Patterns (CAIP'93)*, pages 157–165, Budapest, Hungary, September 1993.

[SLL00] J. Serrat, A. López, and D. Lloret. On ridges and valleys. In *Proc. of International Conference on Pattern Recognition (ICPR'00)*, volume 4, pages 4059–4066, Barcelona, September 2000.

[SM97] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 5(19) :530–535, May 1997.

[SMD+05] A. Shokoufandeh, D. Macrini, S. J. Dickinson, K. Siddiqi, and S. W. Zucker. Indexing hierarchical structures using graph spectral. *Transactions on Pattern Recognition and Machine Intelligence*, 27(7) :1125–1140, 2005.

[SMX04] Z. Sanyuan, Z. Mingli, and Y. Xiuzi. Car plate character extraction under complicated enviroment. In *Proc. of IEEE International Conference on Systems, Man and Cybernitics*, pages 4722–4726, 2004.

[SSDZ98] Kaleem Siddiqi, Ali Shokoufandeh, Sven J. Dickinson, and Steven W. Zucker. Shock graphs and shape matching. In *ICCV*, pages 222–229, 1998.

[Ste96] C. Steger. An unbiased detector of curvilinear structures. Technical report, Technische Universitat Munchen, July 1996.

[Sum05] B. Sumengen. Variational image segmentation demo. http ://aa-kash.ece.ucsb.edu/imdiffuse/segment.aspx, 2005.

[SV52] M. De Saint-Venant. Surfaces à plus grande pente constituées sur des lignes courbes. *Bulletin de la soc. philomath. de Paris*, pages 24–30, 1852.

[SVVV00] J. Stoeckel, F. M. Vos, P. H. Vos, and A. M. Vossepoel. Evaluation of ridge extraction methods for portal imaging. In *Proc. of International Conference on Pattern Recognition*, pages 3433–3436, Barcelona, September 2000.

[Swa91] M. J. Swain. Color indexing. *IJVC*, 7(1) :11–32, 1991.

[TC04] J. Thureson and S. Carlsson. Appearance based qualitative image description for object class recognition. In *ECCV04*, pages 518–529, 2004.

[TMF04] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features : efficient boosting procedures for multiclass object detection. In *CVPR'04*, 2004.

[Vaj00] Z. M. Kovács Vajna. A fingerprint verification system based on triangular matching and dynamic time warping. *Transactions on Pattern Analysis and Machine Intelligence*, 22(11) :1266–1276, November 2000.

[VH99] R. C. Veltkham and M. Hagedoorn. State of the art in shape matching. Technical report, Technical report UU-CS-1999-27, Utrecht, 1999.

[VJ01] P. Viola and M. Jones. Rapid object recognition using a boosted cascade of simple features. In *CVPR'01*, 2001.

[VS91] L. Vincent and P. Soille. Watershed in digital spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6) :583–598, 1991.

[VYV98] L. J. Vliet, I. T. Young, and P. W. Verbeek. Recursive gaussian derivative filters. In B.C. Lovell A.K. Jain, S. Venkatesh, editor, *Proc. of 14th Int. Conference on Pattern Recognition (ICPR'98)*, IEEE Computer Society Press, pages 509–514, Brisbane, August, 16-20 1998.

[WCC04] W. Wu, X. Chen, and J. Chang. Incremental detection of text on road signs from video with application to a driving assitant system. *ACM Multimedia*, 2004.

[Wei93] I. Weiss. Geometric invariants and object recognition. *International Journal of Computer Vision*, 10 :207–231, 1993.

[Wit83] A. P. Witkin. Scale-space filtering. In *International joint Conference on Artificial Intelligence*, pages 1019–1022, 1983.

[WL93] J. Wu and J. Leou. New polygonal approximation schemes for object shape representation. *Pattern Recognition*, 26 :471–484, 1993.

[WMR97] V. Wu, R. Manmatha, and E. M. Riseman. Finding text in images. In *Proc. of the second ACM International Conference on Digital Libraries*, pages 3–12, 1997.

[WMR99] V. Wu, R. Manmatha, and E. M. Riseman. Textfinder : An automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11) :1224–1229, November 1999.

[Wol03] C. Wolf. *Détection de textes dans des images issues d'un flux vidéo pour l'indexation sémantique*. PhD thesis, Institut National des Sciences Appliquées de Lyon, 2003.

[WSV99] N. Winters and J. Santos-Victor. Omni-directional visual navigation. In *International Symposium for Intelligent Robotics System*, pages 109–118, 1999.

[WWP00a] M. Weber, M. Welling, and P. Perona. Toward automatic discovery of object categories. In *CVPR'00*, 2000.

[WWP00b] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV'00*, volume 1, pages 18–32, 2000.

[XHZ01] L. Wenyin X.S. Hua and H.J. Zhang. Automatic performance evaluation for video text detection. In *Proc. of International Conference on Document Analysis and Recognition (ICDAR 2001)*, pages 545–550, Seattle, Washington, USA, September 2001.

[Zha01] L. Zhao. *Dressed Human Modeling, Detection, and Part Localisation*. PhD thesis, The Robotics Institute Carnegie Mellon University, 2001.

[ZLK04] J. Zhao, L. Li, and K. C. Keong. A model-based approach for human motion reconstruction from monocular images. In *Proc. of International Conference on Information Technology for Application (ICITA 2004)*, pages 94–99, China, 07-10 January 2004.

[ZN04] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(9) :1208–1221, Sept 2004.

[ZNL01] T. Zhao, R. Nevatia, and F. Lv. Segmentation and tracking of multiple humans in complex situation. December 2001.

[ZR72] C. Zahn and R. Roskies. Fourier descriptors for plane closed curves. *Computer Graphics and Image Processing*, 21 :269–281, 1972.

[Zuc76] S. Zucker. Region growing : Childhood and adolescence. *Computer Graphics and Image Processing*, 21 :269–399, 1976.

[ZY96] S. C. Zhu and A. L. Yuille. Form : A flexible object recognition and modeling system. *International Journal of Computer Vision*, 20(3), 1996.