Recherche de vidéos par le contenu basée sur l'extraction des images clés par mise en correspondance des points d'intérêts et classification des valeurs de répétabilité

Hana Gharbi

UNIVERSITÉ DE TUNIS EL MANAR
FACULTÉ DES SCIENCES MATHÉMATIQUES, PHYSIQUES ET NATURELLLES DE TUNIS
École Doctorale de Mathématiques, Informatique, Sciences et Technologies de la Matière

THÈSE

présentée en vue de l'obtention du

Diplôme de Doctorat en Informatique

par

Mme. Hana GHARBI

intitulée

Recherche de vidéos par le contenu basée sur l'extraction des images clés par mise en correspondance des points d'intérêts et classification des valeurs de répétabilité

Soutenue publiquement le 27 décembre 2018, devant le jury composé de :

Président :	**M. Mohamed Mohsen GAMMOUDI**	Professeur à l'Institut Supérieur des Arts Multimédia de la Manouba
Rapporteurs :	**Mme. Ikram AMOUS BEN AMOR**	Professeur à l'Ecole Nationale d'électronique et des télécommunications de Sfax
	M. Sadok BEN YAHIA	Professeur à la Faculté des Sciences de Tunis
Examinateur :	**M. Walid BARHOUMI**	Maître de conférences à l'Ecole Nationale d'Ingénieurs de Carthage
Directeur de thèse :	**M. Ezzeddine ZAGROUBA**	Professeur à l'Institut Supérieur d'Informatique de l'Ariana

Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance.
SIIVA : Equipe de recherche Systèmes Intelligents d'Imagerie et Vision Artificielle.

Je dédie ce travail à mes chers parents,

à ma petite Emna et à tous ceux qui me sont chers.

REMERCIEMENTS

Au terme de ce travail, j'adresse mes sincères remerciements à toutes les personnes qui ont contribué à sa réalisation et ont permis par leur soutien et leurs conseils, de le mener à bien. Je tiens tout d'abord à adresser mes remerciements les plus distingués à Messieurs les Membres du Jury : Monsieur Mohamed Mohsen GAMMOUDI, professeur à l'Institut Supérieur des Arts Multimédia de la Manouba, de me faire l'honneur de présider le jury de cette thèse ; Madame Ikram AMOUS BEN AMOR , professeur à l'école nationale d'électronique et des télécommunications de Sfax, et Monsieur Sadok BEN YAHIA, professeur à la Faculté des Sciences de Tunis , d'avoir accepté de rapporter ce manuscrit ; Monsieur Walid BARHOUMI, Maître de conférences à l'Ecole Nationale d'Ingénieurs de Carthage pour avoir bien voulu examiner mes travaux de thèse.

Je remercie particulièrement mon directeur de thèse, Monsieur Ezzeddine ZAGROUBA, professeur à l'Institut Supérieur d'Informatique, qui m'a dirigée durant ce travail doctoral, pour sa patience, son assistance, sa disponibilité et ses précieuses recommandations. Grâce à ses qualités humaines et professionnelles, j'ai pu mener ce travail dans les meilleures conditions.

Je remercie également très chaleureusement Monsieur Sahbi BAHROUN, maître assistant à l'Institut Supérieur d'Informatique, pour son co-encadrement et pour son aide dans mes recherches.

Je remercie affectueusement toute ma famille, pour l'amour et le soutien qu'ils

m'ont toujours offert. Un merci particulier à mes parents, ma sœur, mon frère et mon mari, pour leur amour généreux et constant et de m'avoir toujours encouragée.

Mes remerciements sont adressés, de même, à tous mes collègues au sein du laboratoire LIMTIC et en particulier à Mr Mohamed Massaoudi et Mme Sabra Hechmi. Je remercie enfin tous ceux qui m'ont aidée dans l'accomplissement de ce travail.

ABSTRACT

Video summaries construction is a competitive area of research in the content based video retrieval field. The works presented in this thesis lies in this context whose main objective is to describe the videos of the database by a set of representative keyframes. This process aims to facilitate the content-based video retrieval, which is composed of three phases: the description, the indexing and the retrieval. Thus, the extraction of certain global and local features is a primary task for description and indexing. Most of the state of the art methods used global features. In this work we used local features based on interest points which represents discontinuities. In a first step, we proposed a matching method based on the local description arround interest points using the "Local Binnary Pattern" and the geometric invariants. This method showed its robustness against important interest points matching methods in the literature. It was used to extract features during the indexing process and it served us in the next step, which consists on proposing a new method of video keyframes extraction based on local features. This provides the user with a summary containing the most representative objects in the videos in order to facilitate the search in a video database. In this context, we proposed two variants: The first variant is based on the repeatability table. First, a repeatability table was built based on the proposed matching method. This table contains the repeatability values between frames in the video. Subsequently, the classification of the repeatability values based on PCA and HAC allows the selection of the keyframes that are the centers of the clusters. In order to improve this method, we proposed a second variant. In this variant, we chose a candidate set frames from the video based on a windowing rule and then a repeatability graph was constructed. This graph describes the relation-

ship between the candidate frames in terms of repeatability. The classification of this graph using the modularity maximizing principle facilitates the process of obtaining the representative keyframes of the videos. Finally, we defined an evaluation protocol dedicated to the keyframes extraction methods. In addition to the qualitative and quantitative evaluation, this protocol aims to project the results obtained on content based video retrieval system, in order to ensure more the effectiveness of videos description by the keyframes obtained.
Keywords: Content based video retrieval, keyframes extraction, interest points, repeatability, matching.

RÉSUMÉ

La construction des résumés de vidéos est un domaine de recherche compétitif pour la recherche de vidéos par le contenu. C'est dans ce cadre que se s'inscrivent nos travaux de thèse dont l'objectif principal est de décrire les vidéos de la base par un ensemble d'images clés, et ce pour faciliter la recherche de vidéos par le contenu dans les bases de vidéos. Ainsi, l'extraction de certaines caractéristiques globales et locales s'avère une tâche primordiale. Une des méthodes les plus courantes pour l'extraction d'informations locales s'appuie sur l'utilisation des points d'intérêts représentant une discontinuité. Dans une première étape, nous avons proposé une méthode de mise en correspondance de ces points basée sur la description local autour des points d'intérêts par "Local Binnary Pattern" et sur les invariants géométriques. Cette méthode va nous servir dans la prochaine étape qui consiste à proposer une nouvelle méthode d'extraction des images clés pour chaque vidéo. Ceci permet de fournir à l'utilisateur un résumé contenant les images les plus représentatives dans les vidéos afin de lui faciliter la recherche dans une base de vidéos. Dans ce contexte, nous avons proposé deux variantes : La première variante est basée sur la table de répétabilité. Tout d'abord, la table de répétabilité est construite en se basant sur la méthode de mise en correspondance proposée. Cette table contient les valeurs de répétabilité entre les images de la vidéo. Par la suite, la classification des valeurs de répétabilité permet la sélection des images clés qui sont les centres des classes. Dans le but d'améliorer cet algorithme, une deuxième variante a été proposée. Dans cette variante, des images candidates de la vidéo ont été choisies à l'aide d'une technique de fenêtrage puis un graphe de répétabilité a été construit. Ce graphe décrit la relation entre les images candidates en termes de répétabilité. La classification de ce graphe en utilisant le

principe de maximisation de modularité permet l'obtention des images représentatives de la vidéo. Enfin, nous avons défini un protocole d'évaluation dédié aux méthodes d'extraction des images clés. Ce protocole vise en plus de l'évaluation qualitative et quantitative de projeter les résultats obtenus sur le domaine de recherche de vidéos par le contenu pour s'assurer davantage de l'efficacité de la description des vidéos par les images clés obtenues.
Mots-clés : Recherche de vidéos par le contenu, extraction des images clés, points d'intérêts, répétabilité, mise en correspondance.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLES

LISTE DES ACRONYMES

CBVR : Content Based Video Retrieval

SIFT : Scale-invariant feature transform

SURF : Speeded Up Robust Features

ASIFT : Affine Scale-invariant feature transform

LBP : Local Binary Pattern

CAH : Classification Ascendante Hierarchique

ACP : Analyse en Composantes Principales

HOG : Histogram of Oriented Gradient

GLOG : Gradient Location Orientation Histogram

CBAC : Content based adaptive clustering

RS : Résumé Statique

OVP : Open Video Project

HSV : Hue/Saturation/Value

RGB : Red/Green/Blue

SVD : Décomposition en valeurs singulière

CS : Ensemble des images candidates

CR : Taux de compression

PSNR : Rapport signal bruit

CUS : Comparaison Résumés d'Utilisateurs

LFT : Local Feature Transform

INTRODUCTION GÉNÉRALE

Contexte et motivations

Le développement rapide des technologies multimédia, l'augmentation significative des performances informatiques et la croissance de l'internet ont permis aux utilisateurs d'accéder à un grand volume de données vidéo. Les applications à base vidéos tels que youtube dailymotion...etc sont de nos jours en forte croissance. Ce besoin en croissance exponentielle a boosté la recherche dans le domaine de l'analyse synthèse et indexation de vidéos. Ceci dans l'objectif de fournir des techniques de plus en plus efficaces de fouille dans ce type de bases. Un grand travail de recherche a été fait sur la recherche des vidéos par le contenu au cours des dernières années afin de satisfaire des utilisateurs de plus en plus exigents. Les bases contenant des vidéos sont caractérisées par un volume de données très grand et même gigantesque. Généralement, une seule minute d'une séquence d'un film est équivalente à 1500 images (à une fréquence de 25 images par seconde). Lorsqu'un utilisateur a besoin d'une information spécifique, la tâche de recherche manuelle est difficile et même impossible surtout avec la contrainte de temps. Ceci a poussé les chercheurs à automatiser le processus de recherche et permettre à l'utilisateur l'accès au contenu des vidéos à partir d'une représentation compacte et structurée de ces données. La meilleure représentation était les résumés de vidéos. Ces résumés sont composés d'un ensemble d'images qui doivent contenir les informations les plus pertinentes et les plus représentatives sur les vidéos tout en réduisant la redondance.

Problématique et objectifs

Les résumés de vidéos jouent un rôle essentiel dans les processus d'indexation vidéo, de recherche et de navigation. Plusieurs méthodes de construction de résumé de vidéos ont été proposées dans la littérature. Ces méthodes peuvent être effectuées selon deux alternatives. Le but principal de la première alternative est généralement de donner une vue d'ensemble de toute la vidéo dans le but de donner à l'utilisateur une idée sur le contenu afin de faciliter la recherche et minimiser la complexité du traitement. Pour la deuxième alternative, le résumé permet de prendre en considération l'intérêt de l'utilisateur et les critères à inférer. C'est un résumé qui dépend essentiellement de ce que l'utilisateur souhaite voir (il pourrait par exemple demander un résumé contenant des objets qu'il choisit et non pas tous les objets existants dans la vidéo). Dans ce travail de thèse, nous allons nous intéresser à la première alternative qui est dédiée essentiellement aux applications grand public et qui minimise les tâches que l'utilisateur doit effectuer. Les principaux défis qui doivent être pris en considération lors de la construction du résumé vidéo sont : la précision, la compacité et la minimisation de la redondance.
Plus précisément, notre objectif dans ce travail de recherche est d'aborder le problème de construction de résumé automatique de vidéos tout en prenant en considération les deux caractéristiques principales :

- Le résumé généré doit être le plus possible fidèle au contenu de la vidéo. Toutes les informations pertinentes doivent être présentes dans le résumé avec un minimum d'omission et de redondance.

- Suivre un protocole d'évaluation qui doit prouver la qualité des résumés produits.

- Les résumés générés seront un point de départ pour la recherche par le contenu dans une base de vidéos. L'utilisateur soumettra en requête une image, la

recherche va se faire dans l'ensemble des résumés générés et le système doit retourner l'ensemble de vidéos où l'information cherchée se trouve (visage par exemple).

La plupart des applications de recherche par le contenu y compris la construction des résumés vidéo s'appuient sur l'extraction d'un ensemble de caractéristiques globale ou locale de l'image (que ce soit pour la description des images clés lors du processus d'indexation ou encore pour la construction des résumés). De façon générale, la première étape permet l'identification et la localisation des primitives présentant des propriétés globales ou locales. La majorité des travaux de construction de résumé existants se basent essentiellement sur l'extraction des primitives globales.

Dans le présent travail, nous nous intéressons principalement à l'extraction de primitives locales par extraction et description des points d'intérêts. Les points d'intérêts sont des primitives locales définissant une double discontinuité de la fonction d'intensité dans une image. Après avoir extrait les points d'intérêts, ils seront en second lieu décrits. Cette description est le plus souvent basée sur la relation du point d'intérêt avec son voisinage local (génération de forme géométrique autour des points d'intérêts définissant une région d'intérêt). Ceci permet de faciliter l'étape de mise en correspondance des points d'intérêts entre des couples d'images prises dans des conditions différentes. Les différentes contraintes qu'il faut prendre en considération lors du processus d'extraction de caractéristiques sont essentiellement liées avec le mouvement de caméra. C'est pour cette raison que les primitives détectées et par la suite les descripteurs doivent fournir une robustesse aux différentes transformations de l'image (rotations, changements d'échelle, de point de vue, de luminosité, ... etc).

Tout d'abord, nous avons proposé une première méthode d'extraction des caractéristiques locales. Cette méthode permet d'adapter un descripteur global (le descripteur de texture "Local Binnary Pattern") au contexte des points d'intérêts. Par la suite, nous avons proposé une méthode efficace de mise en correspondance de points d'intérêts

entre des images ayant subi différentes transformations. D'après la littérature, tous les types d'applications telles que la recherche, la reconstruction 3D et la localisation se reposent sur la qualité des points appariés plutôt que leur nombre. Ainsi, dans ce travail, nous nous focalisons sur la favorisation de la qualité des primitives appariées indépendamment de leur nombre. Il existe un grand nombre de travaux pouvant être utilisés afin de résoudre cette problématique. Dans le souci d'apporter une nouvelle proposition qui permet d'améliorer les travaux existants, nous commençons par analyser et comparer les différentes méthodes de mise en correspondance existantes. Puis proposer une nouvelle méthode de mise en correspondance des points d'intérêts basée sur la description locale autour du point d'intérêt et sur les invariants géométriques. Afin de valider la méthode proposée, nous avons établi une comparaison des résultats obtenus avec les résultats des méthodes les plus référencées dans la littérature. Pour ce faire, nous nous sommes basés sur les métriques d'évaluations suivantes:

- La précision, permettant d'évaluer la qualité de la mise en correspondance
- Le taux d'appariement

En raison des objectifs désirés, pour le système de recherche, y compris la construction d'un résumé fidèle au contenu de la vidéo, il est indispensable de présenter un maximum de précision avec un taux d'appariements correct. En effet, une meilleure précision favorise la robustesse aux différentes transformations ou perturbations qui peuvent influencer les images de vidéos.

Nous avons proposé par la suite une première méthode d'extraction des images clés. Cette méthode est basée sur la mesure de répétabilité basée sur la méthode de mise en correspondance déjà proposée. La mesure de répétabilité est couramment utilisée dans la littérature pour juger la ressemblance entre deux images. Nous construirons ainsi une table de répétabilité par plan de vidéo. La valeur (i,j) de cette table est la valeur de répétabilité entre les deux images i et j appartenant au même plan. La

classification de cette table nous a permis de sélectionner les images clés appartenant à chaque plan puis nous avons appliqué le même principe pour toute la vidéo.
Enfin, dans le but d'améliorer la complexité et le temps de calcul, nous avons proposé une deuxième méthode d'extraction d'images clés où le traitement s'effectue seulement sur une ensemble d'images sélectionnées par la technique de fenêtrage et non pas sur la totalité des images de la séquence vidéo comme la première méthode proposée. Nous avons introduit la notion de graphe pour faciliter la sélection des images clés. Cette méthode a montré un compromis entre les résultats retournés, la quantité en terme de redondance et la complexité. Ainsi, pour montrer davantage son efficacité, nous l'avons testée dans un système de recherche de vidéos par le contenu. Les résultats confirment que les méthodes proposées pour le résumé vidéo ont permis une recherche pertinente dans une base de vidéo.

Structure du manuscrit

Le présent manuscrit est divisé en trois chapitres qui sont organisés comme suit:
Le premier chapitre pose le cadre théorique de nos travaux. Nous présentons le contexte de recherche de vidéo par le contenu ainsi l'une de ses étapes de base qui est la construction de résumé des vidéos. Afin de situer nos travaux par rapport à leur contexte, nous avons introduit le processus d'extraction des primitives et nous nous sommes focalisés sur le processus général d'extraction des descripteurs locaux ainsi que ses différentes étapes.
Le deuxième chapitre présente les différentes contributions proposées dans ce travail de thèse. Nous avons commencé par présenter la méthode de mise en correspondance proposée. Par la suite, nous avons présenté les différentes variantes de notre méthode proposée pour la construction de résumé sous forme d'images clés.
Dans le troisième chapitre, nous avons effectué une étude expérimentale qui montre des exemples de résultats obtenus pour chacune des méthodes proposées. Par la suite,

nous avons effectué une étude comparative avec un ensemble de méthodes les plus citées dans la littérature et ce afin de montrer l'efficacité de la méthode proposée. Pour s'assurer davantage de l'efficacité de la méthode proposée, nous avons projeté nos résultats dans un protype de système de recherche de vidéos par le contenu. Nous terminerons ce manuscrit par une conclusion générale et des perspectives.

Chapitre 1

RECHERCHE DE VIDÉOS PAR LE CONTENU : ÉTAT DE L'ART

Introduction

Nous présentons dans ce chapitre le contexte général dans lequel s'inscrit ce travail, ainsi que les différents problèmes abordés aux cours de cette thèse. Nous exposons en premier lieu la problématique concernant les systèmes de recherche et d'indexation de vidéos par le contenu. Nous passons en revue l'historique des différentes méthodes de résumé de vidéos présentées dans la littérature. Une attention particulière sera accordée aux résumés statiques issus d'une sélection d'images représentatives appelées images clés et plus précisément celles basées sur la description locale par points d'intérêts. Nous justifierons ce choix par rapport aux autres méthodes existantes. Nous concluons chaque partie de ce chapitre par une discussion.

1.1 Indexation et recherche des vidéos par le contenu

Les systèmes de recherche de vidéos par le contenu CBVR (Content Based Video Retrieval) se basent essentiellement sur trois phases : la phase de description, la phase d'indexation et celle de recherche. Une étape de prétraitement sera ajoutée ou la vidéo est représentée par un ensemble d'images clés décrivant les informations les plus pertinentes de chaque vidéo. Dans cette partie, nous allons décrire le principe général de chacune de ces phases. Nous présentons dans la figure 1.1 l'architecture générale du système de recherche de vidéo par le contenu.

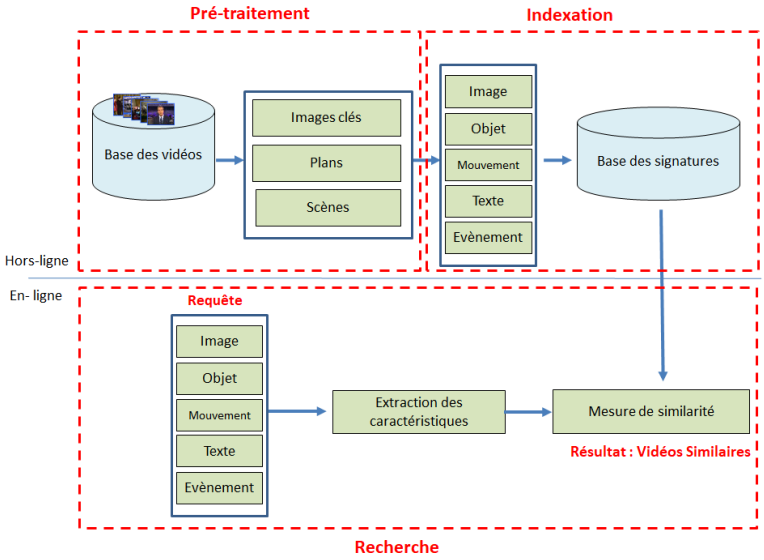

Figure 1.1. Architecture générale du système de recherche de vidéos par le contenu

1.1.1 Phase de description

La description de vidéo, considérée comme étape de prétraitement, permet de faciliter la recherche des informations contenues dans une base composée d'un nombre important de vidéos. En effet, la description de vidéo doit répondre à certains critères nécessaires qui sont : la visualisation rapide du contenu des vidéos et la manipulation facile des informations générées en résultat. Parmi les solutions importantes dans la littérature pour la description de vidéo on peut trouver les résumés statiques qui permettent la représentation de vidéos contenues dans une base de vidéo par un ensemble des images clés. Ces images clés doivent décrire fidèlement le contenu des vidéos.

1.1.2 Phase d'indexation

La phase d'indexation, se déroulant elle aussi en mode hors ligne, consiste à organiser les documents existants dans une base de données. Ceci permet d'extraire les informations nécessaires qui facilitent la représentation et l'organisation des éléments de la base afin de faciliter la sélection sous certains critères. Pour ce faire, il faut commencer par traiter les documents pour extraire les primitives nécessaire tel que : couleurs, textures, objet pertinents, visages. Par la suite, représenter ces informations sous forme de signatures numériques pour ne conserver que le nécessaire et éviter un grand volume de données à stocker. Comme une dernière étape, ces signatures sont structurées pour faciliter la recherche et l'accès aux informations désirées.

1.1.3 Phase de recherche

C'est dans cette phase où l'utilisateur doit introduire sa requête. Cette requête est un peu plus difficile à exprimer dans les images et les vidéos que dans le texte. Cependant, elle peut être sous différentes formes (objets, images, texte,.). La formulation correcte de la requête est l'étape clé pour la réussite de la recherche. Le système s'occupe de transformer la requête en une signature afin de pouvoir la comparer avec celles existantes dans la base et trouver les informations désirées. Cette étape se déroule en ligne. Vu la difficulté de répondre aux besoins de l'utilisateur à partir de sa première requête, il est toujours possible d'ajouter une phase supplémentaire de bouclage de pertinence (BP) comportant la possibilité à l'utilisateur pour modifier sa requête en prenant en considération les résultats déjà obtenus et obtenir des résultats meilleurs.

1.2 Description de vidéos : Résumé vidéo

En comparaison avec le texte et l'audio, la vidéo est le type de données multimédia préféré pour l'être humain. Ceci en raison de la quantité d'informations abondantes qu'elle peut fournir. Avec le progrès rapide des technologies informatiques et réseaux,

la quantité de données vidéo augmente rapidement, ce qui entraîne inévitablement un cout spatio-temporel énorme. Le défi consiste à gérer, classer et récupérer les données massives des vidéos et puis la récupération de leurs contenus, d'où l'apparition des technologies de recherche de vidéos basée sur le contenu (CBVR). Le système de recherche de vidéo basée sur le contenu comprend plusieurs technologies comme la détection des plans, le résumé de vidéo, l'analyse des scènes, etc. Le résumé joue un rôle essentiel pour indexation vidéo, la recherche et la navigation. D'après la littérature, on peut distinguer deux types de résumés : statique et dynamique. Dans ce manuscrit, on va se focaliser sur le résumé statique qui génère des images appelées images clés et qui se présente sous forme de connexion entre l'étape de détection de plans et l'étape d'acquisition des informations sémantiques avancées.

1.2.1 Structure des vidéos

La vidéo est un ensemble d'images regroupées et affichées à environ (25 ou 30) images par seconde. Dans le contexte de notre thèse, uniquement l'information visuelle a été considérée (le son n'a pas été étudié). La structure de vidéo se compose de plusieurs niveaux d'abstractions. Le niveau de base pour les vidéos est appelé "plan". On peut le définir par un ensemble d'images successives dans la vidéo filmée de façon continue sans aucun effet spécial et aucune coupure. A partir du plan, plusieurs niveaux ont été définis suivant le regroupement des images. Le niveau suivant des plans sont les scènes (appelés aussi macro-segments). Ce niveau consiste à regrouper les plans selon un critère bien défini tel que par exemple la similarité des contenus visuels, la similarité des mouvements ou la similarité sémantique en personne ou objet. La Figure 1.2 montre la structure hiérarchique générique d'une séquence vidéo dans le processus d'extraction des images clés.

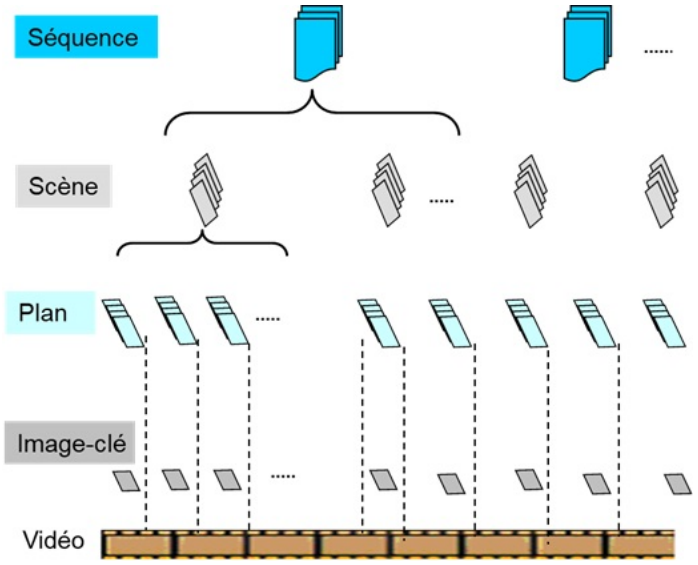

Figure 1.2. Structure hiérarchique générique d'une séquence vidéo

1.2.2 Détection de Plans

La segmentation de vidéo en plans est la première étape dans le processus d'indexation de vidéos numériques pour un système de navigation et de recherche réussi. Dans le cas de génération de résumé, elle est considérée comme étape de prétraitement. Selon la littérature, on trouve plusieurs techniques de détection de plans de la vidéo. Ces techniques peuvent être classées en différentes catégories. Selon [SenGupta, 2015], les principales catégories de techniques de détection de plans de vidéos sont: celles basées sur la comparaison de pixels ou des blocs, celles basées sur la comparaison des histogrammes (globales et locales), celles basées sur la classification et celles basées sur les caractéristiques (locales et globales). Nous présentons dans la figure 1.3 la structure hiérarchique des techniques de détection des plans de vidéo.

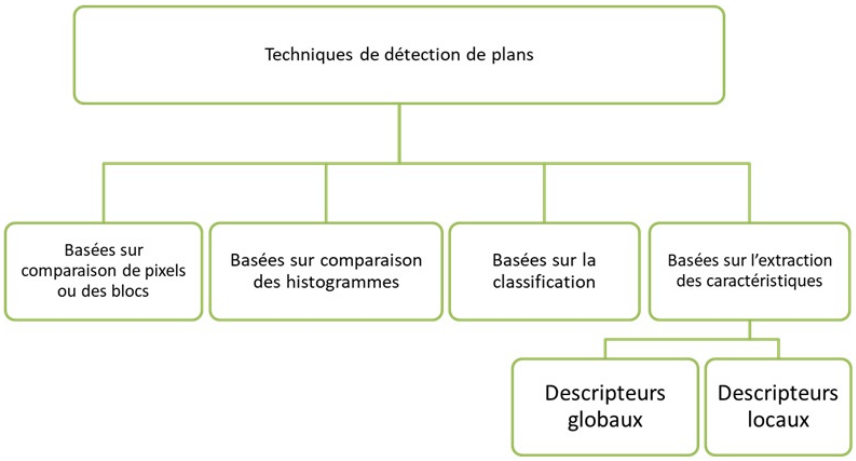

Figure 1.3. Structure hiérarchique des techniques de détection de plans [Sen-Gupta, 2015]

a - Techniques basées sur la comparaison de pixels ou des blocs

La comparaison des paires de pixels entre deux images consécutives permet d'évaluer la différence en intensité du pixel correspondant.

$$D(f\,,\,f+1) = \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y}(I_f(x,y) - I_{f+1}(x,y))}{X,Y} \tag{1.1}$$

avec f, $f+1$ sont deux images adjacentes de taille X x Y, I_f est la valeur d'intensité de pixel de coordonnées (x, y). Il est clair que cette méthode n'est pas efficace car un simple changement de caméra ou mouvement d'objet peut donner une large différence des intensités des pixels. Concernant les méthodes basées sur la comparaison des blocs, leur principe consiste à diviser les images de la vidéo en blocs. Puis, chaque bloc est comparé avec son correspondant dans les images consécutives. Une transition est signalée lorsque le nombre de blocs qui changent est supérieur à un seuil prédéfini.

b - Techniques basées sur la comparaison des histogrammes

Les histogrammes, représentent le nombre de pixels qui ont une couleur fixe dans chaque rang de couleur, ils peuvent être construits dans plusieurs types d'espaces couleurs tels que RGB, HSV, CMYK, YC_bC_r,etc. Ainsi, la comparaison peut être exécutée de façon globale en comparant les histogrammes de chacune des deux images consécutives et si la différence est supérieure à un seuil une transition sera signalée.

$$D(f\ ,\ f+1) = \sum_{i=1}^{n} \mid H_f(i) - H_{f+1}(i) \mid \tag{1.2}$$

Où $H_f(i)$ est la valeur d'histogramme pour le niveau de gris ou couleur i d'une image f et n est le nombre total de niveaux de gris. Localement, on peut diviser les images en blocs puis la comparaison se fait sur les histogrammes de blocs.

$$D(f\ ,\ f+1) = \sum_{k=1}^{x} \sum_{i=1}^{n} (H_f(i,k) - H_{f+1}(i,k)) \tag{1.3}$$

Où $H_f(i,k)$ est la valeur d'histogramme d'une couleur i pour le bloc k d'une image f et x est le nombre total de blocs.

c - Techniques basées sur la classification

Le principe des méthodes de détection de plans par classification consiste initialement à sélectionner au hasard n images comme centres de classes. Les distances entre chaque image et les centres des classes sont calculées et les images ayant une plus petite distance sont classifiées ensemble. Parmi les algorithmes de classification qui ont été largement utilisés pour la détection de plans on peut citer : la classification floue [Chun, 2001], moyenne sift [Lu, 2010], etc.

d - Techniques basées sur la description des caractéristiques

Généralement, ces techniques se basent sur le calcul des descripteurs globaux (couleur, texture, motion, contour,..) [Thounaojam, 2014] pour chaque image puis faire la

comparaison avec l'image suivante en comparant les signatures. Les méthodes les plus simples sont basées sur le calcul de nombre de pixels de contour entre les images consécutives peut donner une information sur les bords du plan [Zabih, 1999]. Un autre exemple est proposé dans [Zhao, 2008] basé sur le calcul d'entropie conjointe et l'information mutuelle entre chaque couple d'images successives puis les bords des plans sont déterminés en utilisant "Canny edge detector".

Il existe aussi quelques techniques de détection de plans basées sur des descripteurs locaux tel que par exemple : SIFT [Lowe, 2004] et SURF [Bay, 2008], etc. Dans cette famille de techniques, les descripteurs locaux des images successives sont calculés puis comparés. Lorsque le nombre de points d'intérêts appariés est inférieur à un seuil prédéfini, un nouveau plan est détecté. Parmi les travaux se basant sur la description locale dans la littérature, on peut citer celui de Deepak et al, [Deepak, 2013] qui utilise le corrélograme de couleurs et le descripteur des points d'intérêts G-SURF "Gauge Speed up robust feature" [Alcantarilla, 2013]. Initialement, ils ont conçu une application linéaire de corrélograme de couleurs et G-SURF pour détecter les transitions de plans. Ensuite, pour améliorer les performances cette méthode est utilisée dans l'extraction des images clés. Dans le travail proposé par [Shekar, 2011], le vecteur descripteur de chaque image de la vidéo est calculé en utilisant la transformation des descripteurs locaux (LFT : local feature transform). Après avoir appliqué le LFT pour chaque image, on extrait les descripteurs puis chaque premier et deuxième moment sont calculés pour les espaces de couleurs des canaux afin de calculer les vecteurs caractéristiques. Dans une autre alternative, une image est considérée comme bord du plan si elle ne contient pas (ou contient juste un peu) de points appariés avec les images qui la suivent. Pour plus de robustesse, l'entropie est calculée pour toutes les images puis chacune est comparée avec celles adjacentes [Baber, 2013]. Cette méthode de segmentation de vidéos combine entre la description locale (SURF) et globale (entropie) pour la segmentation des vidéos.

1.2.3 Construction des résumés vidéos

Une fois les méthodes de détection de plans décrites, nous allons passer à la description des familles des techniques de résumé vidéo. Deux grandes familles de méthodes de construction de résumé se distinguent dans la littérature: le résumé statique et le résumé dynamique. Ces deux familles vont être présentées puis discutées dans les paragraphes suivants.

1.2.3.1 Résumé statique

Le résumé statique de vidéo se compose d'un ensemble d'images qui doivent représenter le contenu de la vidéo. Chacune de ces images est soigneusement extraite pour qu'elle puisse représenter le contenu visuel de chaque partie de la vidéo sans redondance. Le résultat de ce type de résumé est facile à visualiser et il minimise la complexité du processus de recherche. Beaucoup de travaux ont été réalisés ces dernières années et trois familles principales de résumé statique ont été dégagées : méthodes reposant sur l'échantillonnage, la classification, l'extraction des caractéristiques et autres. Nous avons repris ces trois familles de résumé que nous avons complétées avec d'autres travaux qui sont soit basés sur la combinaison des catégories précédentes, soit sur d'autres alternatives. La façon formelle de représenter un résumé statique d'une séquence vidéo peut être définie comme suit :

$$RS(S) = \{Image_1, Image_2, Image_3, ..., Image_N\} \tag{1.4}$$

Où RS est le résumé statique, la séquence vidéo est S, $Image_i$ relative à la ième image clé extraite de S et N est le nombre de toutes les images représentatives incluses dans le résumé. La qualité du résumé résultant est influencée essentiellement par le paramètre N, d'où la nécessité de résoudre le problème lié à la valeur de N. La première solution était de proposer une méthode qui consiste à fixer à priori le nombre d'images clés à considérer. Cependant, ce n'est pas facile de fixer une seule valeur et imaginer qu'elle pourrait être adaptée et valable pour n'importe quel type de vidéo

rencontrée en pratique. Ainsi, la façon idéale est que les méthodes de résumé vidéo peuvent être en mesure de déterminer de manière adaptative les valeurs adéquates du paramètre N. Nous présentons dans la figure 1.4 la structure hiérarchique des méthodes de résumé statique de vidéo.

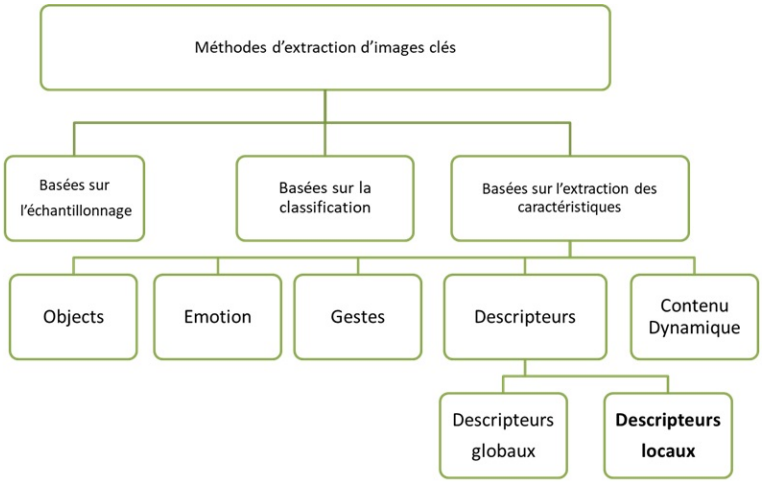

Figure 1.4. Structure hiérarchique des méthodes de résumé vidéo

a - Méthodes basées sur l'échantillonnage

Les premiers travaux d'extraction des images clés commencent par proposer des méthodes très naïves telles que le sous-échantillonnage de vidéo (ou plans de vidéo) uniformément et de façon aléatoire. Ils choisissent par exemple pour chaque vidéo : la première image, la dernière ou celle du milieu [Tonomura, 93], comme le montre la figure 1.5. L'inconvénient majeur de ces méthodes est qu'ils ne prennent pas en considération le contenu des images, ce qui peut donner des images clés avec un contenu soit redondant soit avec informations manquantes.

D'autres méthodes utilisent la technique d'échantillonnage après une étape de détection de plans. Ceci dans l'objectif d'assurer l'obtention des images clés adaptées au contenu de la vidéo. Parmi les méthodes les plus simples qui existent dans ce contexte,

on peut citer celle proposée par [Nagasaka, 91] qui considère que la première image de chaque plan comme image clé, et celle proposée par [Ueda, 91] qui considère la première et la dernière image de chaque plan. Ces méthodes ne sont efficaces que pour des vidéos où il n'existe pas de grande variation dans le contenu ou des variations importantes des caméras. Dans [Pentland, 94], la sélection des images clés consiste à fixer des intervalles tout au long de chaque plan, puis prendre des endroits prédéfinis comme images clés. Dans les travaux de [Rui, 98], les images clés seront le résultat d'un échantillonnage du plan selon des intervalles prédéfinis de temps.

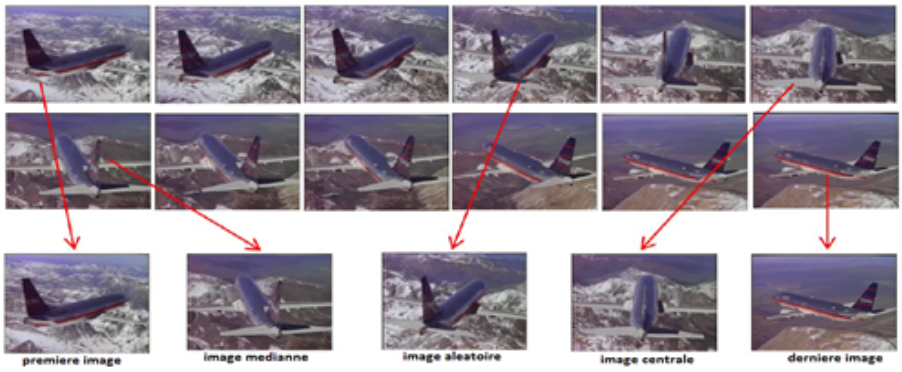

Figure 1.5. Résumé statique basé sur l'échantillonnage

b - Méthodes basées sur l'extraction des caractéristiques

Ils existent plusieurs méthodes de construction de résumé, dans la littérature, basées sur l'extraction des caractéristiques. Les méthodes de construction des résumés peuvent être globales en utilisant des caractéristiques globales (exemples : mouvement, couleurs, texture, gestes, objets, contenu dynamique,..). Elles peuvent aussi être locales, se basant sur des caractéristiques locales tels que les régions et les points d'intérêts.

Dans les exemples de [Zhang, 95] et [Gunsel 97], les auteurs proposent de commencer par sélectionner la première image de chaque plan comme image clé puis en se basant

sur un seuil, ils sélectionnent le reste des images clés. Le seuil prédéfini est basé sur la distance entre les histogrammes des images de chaque plan. Ces méthodes ne sont pas pertinentes vu qu'elles sont fortement dépendantes des seuils et des effets de transition des plans. Dans [Ciocca, 06], un autre exemple a été proposé. Cette méthode est basée sur une courbe caractérisant les images de vidéos selon leurs descripteurs : des histogrammes d'orientation et de couleur et des ondelettes. Ainsi, les images clés seront les points appropriés sur la courbe. Dans [Zhonghua, 2004], les images du plan sont segmentées en objets et arrière-plan. Puis, un calcul du rapport entre les objets et arrière-plan sera effectué. Les images clés sont celles qui possèdent un rapport maximum. Les auteurs considèrent que ces images donnent plus d'informations sur chaque plan que les autres.

Dans Chergui et al [Chergui, 2012], une méthode basée sur la description locale a été proposée. La méthode proposée suppose que la vidéo est déjà segmentée en plans. L'image contenant un nombre maximal de points d'intérêts, dans chaque plan, est considérée comme la plus riche en informations. Ainsi, elle sera sélectionnée comme image clé et représentera le contenu de ce plan. Or, ce n'est pas toujours valable car le nombre de points d'intérêts existants dans les images de la vidéo n'a aucune relation avec son contenu. Aussi, une image par plan n'est pas toujours valable et ça peut causer soit une perte d'informations lorsque le plan est riche en contenu, soit une redondance lors de la répétition d'un ou de plusieurs plans dans la vidéo. Dans une autre méthode proposée par Tapu et al. [Tapu, 2011] le nombre d'images clés n'est pas fixé à priori. Par contre un paramètre N est fixé. Puis, la première image clé sélectionnée est l'image numéro N dans chaque plan. D'autres images aux positions multiples de N seront sélectionnées pour être analysées. Chacune des images sélectionnées est comparée avec l'image clé extraite précédemment. Si la dissimilarité visuelle calculée à travers la distance "chi-square" des histogrammes couleurs HSV est élevée, l'image en cours sera ajoutée à la liste des images clés. Finalement, ils éliminent les images clés qui ne sont pas importantes en se basant sur le nombre

de points d'intérêts détectés par le détecteur SIFT. Dans [Li, 2012], les auteurs ont définit les changements de contenu et de complexité de vidéos en utilisant le détecteur SIFT. Puis, en se basant sur le résultat, ils fusionnent les plans similaires. Le nombre d'images clés sera le nombre de plans après la fusion. Cette méthode donne un bon résultat grâce à la robustesse du détecteur SIFT contre différentes variations. Mais, son utilisation cause une complexité de calcul élevée. Dans [Barbieri, 2014], une autre technique utilisant le descripteur SIFT a été proposée. Les auteurs commencent par générer un nombre d'images candidates puis éliminer celles qui se ressemblent en définissant un pourcentage de différence de nombre de points d'intérêts.

c - Méthodes basées sur la classification

D'autres travaux ont été réalisés en utilisant le principe de la classification dans la sélection des images clés. Les chercheurs considèrent que la classification permet d'obtenir des images représentatives du contenu des vidéos qui sont différentes entre elles. Ainsi, ils proposent une classification globale de toute la vidéo. Le fait de choisir une image par classe pour être insérer dans le résumé minimise la redondance. Ils utilisent des caractéristiques différentes (les descripteurs, les mouvements, les histogrammes,..) et suivent plusieurs méthodes de classification (k-moyenne, c-moyenne, classification hiérarchique,..).

Parmi les méthodes proposées basées sur la classification, on peut citer celle de Ghirgenshn et al [Girgensohn , 2000]. Cette méthode utilise la classification hiérarchique afin d'obtenir des résumés dans plusieurs niveaux d'abstractions puis impose des contraintes de temps sur la position des images clés. Sun et al [Sun, 2000] proposent un algorithme qui utilise CBAC "Content based adaptative clustering" dont lequel les changements de contenu sont classés dans l'ordre croissant puis comparés pour déterminer les parties les plus importantes. Gong et al [Gong, 2001] ont utilisé la décomposition en valeurs singulières SVD pour extraire les résumés. Dans cette méthode chaque image sera représentée par un vecteur caractéristique. En effet, elle

sera divisée en 9 régions qui seront représentées chacune par un histogramme couleur. Ensuite, les histogrammes seront couplés pour former le vecteur caractéristique. La SVD sera par la suite appliquée sur la matrice composée de vecteurs caractéristiques. Le résumé sera composé selon la taille choisi par l'utilisateur.

Pour Asadi et al. [Assadi, 2012], la première étape consiste à segmenter la vidéo en plans puis prendre des images situées dans des intervalles échantillonnés pour extraire les informations nécessaires pour obtenir leurs histogrammes couleurs. A la fin, ils groupent les images en utilisant "fuzzy c-mean clustering", sélectionne une image pour chaque classe et les organisent selon leur ordre chronologique. L'avantage de cette méthode est que la matrice résultante pour chaque classe donne une information sur l'image la plus informative. Dans [Hannane, 2016], l'approche proposée consiste à appliquer l'SVD "Décomposition en valeurs singulières" pour toutes les images du plan puis calcule l'entropie à travers le vecteur résultant de l'application de l'SVD. L'image clé relative à chaque plan est celle qui possède une large valeur d'entropie.

d - Autres méthodes

Chacune des catégories d'extraction d'images clés présentées précédemment a ses avantages et ses inconvénients. Par exemple les méthodes basées sur un échantillonnage de la vidéo ou des différents plans sont moins coûteuses en temps de calcul et de complexité par rapport aux méthodes basées sur les caractéristiques et surtout sur la classification. Par contre, ces dernières sont plus robustes et minimisent considérablement la redondance ou paradoxalement le temps de calcul.

C'est pour cette raison, plusieurs auteurs ont eu recours à la combinaison de plusieurs catégories ou à autres alternatives. Dans ce contexte, on peut citer la méthode proposée par N.Ejaz et al. [Ejaz, 2012] dans laquelle les auteurs ont proposé un mécanisme d'agrégation dans le but de combiner les caractéristiques visuels extraites à partir de la corrélation des canaux de couleurs RGB, des histogrammes couleurs et des moments d'inertie pour extraire les images clé. Dans [Sandra, 2011], les au-

teurs proposent une méthode basée sur la combinaison des descripteurs couleurs et d'algorithme de classification k-moyenne. Cet algorithme donne des bons résultats en termes de redondance mais il est coûteux en termes de complexité. J.L.Lai, et al [Lai, 2012] ont utilisé un modèle d'attention visuel basé sur la mesure de saillance et ont sélectionné les images ayant une valeur de saillance maximale comme image clé. Dans M. Kumar et al [Kumar, 2011], les auteurs ont analysé les informations spatio-temporelles de la vidéo pour une représentation éparse, ils ont utilisé la classification normalisée pour générer les classes et l'image milieu de chaque classe ordonnée temporellement est sélectionnée comme image clé. Sergent et al. [Sergent, 2015] proposent une méthode de résumé plus évoluée basée sur le traitement analytique en ligne de données. Une telle structure intègre différents outils tel que l'opération de driller qui permet de parcourir efficacement les descripteurs multiples de bases de données en fonction de niveau de détail élevé.

1.2.3.2 Résumé dynamique

Passons maintenant au résumé dynamique. Il est considéré meilleur à interpréter que le résumé statique vu qu'il garde l'aspect dynamique de la vidéo ainsi que l'information audio-visuelle. Il est donc plus proche de la vidéo originale. Cependant, les méthodes qui génèrent ce type de résumé sont plus sophistiquées et plus coûteuses en terme de temps et mémoire de stockage nécessaires. D'après la littérature [Boukadida, 2015], ces méthodes sont principalement regroupées en quatre catégories :

a - Méthodes basées sur les modèles d'attention

L'objectif de ces méthodes est de simuler et de modéliser l'attention des utilisateurs pour créer le résumé vidéo. Exemple : [Gygli, 2014], [Longfei, 2008], [Li, 2009], etc. Ces méthodes procèdent généralement en trois étapes:
La première étape consiste à segmenter toute la vidéo en plusieurs unités de base.

L'unité de base peut être soit temporelle (seconde, minute, ...), soit sous forme de segments (scènes, plan, ...) ou même des images.

La deuxième étape, consiste à calculer un score pour chacune des unités de base précédemment extraites. Ce score reflète l'importance qui peut être attribuée à l'unité de base et donne une indication sur la probabilité pour qu'elle soit sélectionnée et incluse dans le résumé. En pratique, les scores sont calculés en se basant sur des algorithmes de détection de caractéristiques. Ces caractéristiques permettent d'identifier des moments susceptibles d'attirer l'attention humaine. Les scores calculés sont utilisés par la suite pour créer une courbe modélisant l'attention des utilisateurs. Par exemple, l'attention de l'utilisateur est souvent captée par des événements visuels, acoustiques ou textuels. Par conséquent, des caractéristiques visuelles, acoustiques et textuelles sont détectées pour créer respectivement des courbes d'attention visuelles, acoustiques et textuelles. Une courbe de synthèse modélisant l'attention humaine est obtenue par la fusion de ces différentes courbes.

La troisième étape consiste à analyser la courbe d'attention finale obtenue précédemment et ainsi sélectionner les extraits qui doivent être inclus dans le résumé. Ces extraits correspondent aux pics de la courbe utilisant un seuil ou selon un intervalle de temps (par exemple sélectionner les extraits d'une durée fixée à priori qui ont un pic au centre).

b - Méthodes basées sur la vue d'ensemble

D'autres méthodes visent à générer des résumés sous forme d'aperçus. Exemple : [Li, 2011], [Shroff, 2010], etc. Ces aperçus seront utilisés pour permettre aux utilisateurs de prendre une idée générale sur l'ensemble de la vidéo. La façon la plus simple de créer ce type de résumé est d'accélérer la vidéo. L'idée consiste à condenser la vidéo originale, simplement, en accélérant sa lecture. Bien que ce type de résumé permet de couvrir la totalité du contenu de la vidéo, il ne peut pas nécessairement attirer l'attention des utilisateurs.

c - Méthodes basées sur l'extraction d'événements intéressants

Un autre type de résumé vidéo consiste à détecter les moments forts de la vidéo (highlights). Exemple : [Liu, 2010], [Tsai, 2013], etc. Mais, ceci est considéré compliqué sans une connaissance du type de la vidéo à priori. Ainsi, pour être de plus en plus efficace, les résumés suivant des extraits clés se basent généralement sur des hypothèses très élémentaires.
Dans ce cas, ces méthodes seront ciblées aux vidéos de types spécifiques (sport et émission de journal par exemple) ou les moments forts présentent des caractéristiques particulières. Ces caractéristiques peuvent être utilisées par la suite pour la phase d'apprentissage utilisée généralement pour chaque type de vidéos.

d - Méthodes basées sur les Tweets

Ces méthodes se basent principalement sur les messages courts des microblogues (récupérés à partir des réseaux sociaux comme Twitter par exemple). Exemple : [Tang, 2012], [Takamura, 2011], etc). Le service de microblogage est un service de messages courts qui permet aux utilisateurs de poster brièvement leurs messages. Ces messages courts sont appelés tweets. Le principe générale de ces méthodes permet l'association du contenu textuel avec me contenu audiovisuel dans le but de déterminer les segments les plus importants dans une vidéo.

1.2.3.3 Bilan et discussion

Le choix du type de résumé (dynamique ou statique) dépend fortement du type d'application ainsi qu'aux exigences de l'utilisateur. Dans nos travaux de thèse, nous nous interessons à la recherche de vidéo par le contenu. L'avantage majeur des quatre types de résumé dynamique est qu'ils préservent l'information temporelle des segments de vidéos résultants. Cependant, malgré la conservation de l'aspect temporel et dynamique, le problème linéaire persiste ce qui oblige l'utilisateur à regarder

tout le segment de la vidéo pendant toute la durée prédéfinie afin de prendre une idée générale résumant le contenu de la vidéo. De ce fait, nous avons opté pour la représentation statique du résumé vu qu'elle est moins coûteuse en termes de mémoire, temps de calcul et écarte le problème de linéarité.

Chacune des méthodes de résumé statique possède des avantages et des inconvénients. Bien que les méthodes basées sur l'échantillonnage soient les moins coûteuses en termes de complexité et temps d'exécution, elles présentent un inconvénient majeur. En effet, ces méthodes ne prennent pas en considération le contenu des images de la vidéo. Ce qui peut donner des images clés avec un contenu soit redondant soit avec informations manquantes.

En ce qui concerne les méthodes basées sur la classification, elles sont les plus efficaces que celles basées sue l'échantillonnage car elles prennent en compte le contenu des images. Par contre, elle ont un temps de calcul plus élevé. Ces dernières sont dans la majorité des cas couplées avec des descripteurs globaux. Un descripteur global est calculé sur chaque image du plan. Une classification relationnelle est appliquée par la suite sur les descripteurs globaux afin de déterminer des classes regroupant les images ayant une certaines similarité selon le descripteur global déja calculé. Le centre de classe est pris comme échantillon représentatif de toute la classe et par conséquent, il est considéré comme image clé. Les méthodes d'extraction d'images clés basées sur la description globale ont été largement étudiées dans la littérature. Bien que la description locale a montré son éfficacité et sa robustesse par rapport à la description globale, elle a été très peu utilisée, dans ce contexte.

En effet, l'utilisation des points d'intérêts comme descripteur local pourra être une bonne alternative pour l'extraction d'images clés. Les méthodes existantes dans ce contexte, basée sur cette alternative, ne sont pas robustes en comparaison avec celles utilisant les descripteurs globaux dans la littérature. Ainsi, dans nos travaux de thèse nous allons mettre l'accent sur la caractérisation locale en utilisant les points d'intérêts. Les points d'intérêts ont l'avantage de fournir des descripteurs robustes

contre plusieurs types de transformations tels que les rotations, les changements affines, les changements d'illuminations et les occultations.

1.3 Indexation de vidéos par description locale

La génération de résumé vidéo a pour objectif la présentation des passages qui permettent à l'utilisateur de comprendre au maximum la vidéo. Pour ce faire, il est donc nécessaire d'extraire certaines caractéristiques afin de comprendre et analyser l'aspect sémantique de la vidéo. Malgré les importantes avancées dans le domaine de traitement du contenu d'images et de vidéos, le problème d'extraction automatique de la sémantique reste un challenge. Plusieurs types de descripteurs ont été proposés dans la littérature. Le choix du descripteur reste un défi car cela dépend de plusieurs facteurs tels que l'objectif désiré et le type d'application. Plusieurs types de classifications de descripteurs ont été proposés dans la littérature. De façon générale on trouve deux catégories : les descripteurs globaux et les descripteurs locaux.

La caractérisation en utilisant la description globale discutée dans la section précédente se focalise dans le traitement de l'image de façon globale. Cependant, dans certains cas, on a besoin des informations supplémentaires sur un objet ou une partie bien précise dans l'image. Dans ce cas, les caractéristiques globales peuvent ne pas donner le résultat désiré. On peut citer le cas où on a une image avec des objets différents qui sont chacun décrit avec des caractéristiques globales (couleur, forme, texture...). Dans ce cas, le vecteur caractéristique résultant peut ne pas nous donner l'information précise sur chaque objet localement ainsi que leurs dispositions dans l'image. La segmentation d'images en régions ou blocs peut faciliter l'obtention des informations locales mais cela ne donne pas toujours la performance voulue. Ceci est du aux différentes lacunes des algorithmes de segmentation.

Pour remédier à ce problème, plusieurs travaux dans la littérature ont utilisé les points d'intérêts comme primitive locale[Yuheng, 2017]. Leur utilisation a montré une effi-

cacité dans plusieurs applications telles que la recherche et la reconnaissance d'objets. Dans ce contexte, nous allons citer les deux principales catégories pour la description locale, et justifier notre choix pour la description locale par points d'intérêts.

1.3.1 Méthodes de description locale par régions

Pour les méthodes de description des images par régions la question la plus importante qui se pose : comment les images sont divisées en régions et quel est le descripteur qui permet une comparaison efficace des images entre elles ? Certaines techniques se basent sur la composition en blocs [Safavian, 1997]. Ils consistent à diviser simplement les images en grilles rectangulaires de tailles fixes. Ces techniques sont simples, mais, le découpage des images en blocs n'est pas une solution efficace et ceci pour une simple raison : chaque bloc n'est pas dans la majorité des cas un objet réel ce qui rend le découpage non significatif. La décomposition en régions de tailles fixes ne donne pas satisfaction car la résolution avec laquelle les images sont acquises laisse le choix de la taille de ces blocs non triviale. D'autres techniques se basent sur la segmentation des images en régions [Yuheng, 2017][Huang, 2017]. Cela aura un sens surtout dans le cas où cette segmentation fournit ce qu'on appelle région-objets. Mais on aura toujours des informations manquantes au niveau de la localisation quand l'image est riche en contenu. Les frontières entre les objets sont difficiles à extraire lorsque la scène est encombrée. Une liste de descripteurs des régions des images existe. Ces descripteurs sont généralement inspirés des descripteurs globaux, représentés sous forme de signatures des régions qui seront utilisées pour entamer le processus d'indexation. Parmi les nombreux descripteurs existants on peut citer les suivants :

- Histogramme de couleur qui s'intéresse à la description de la couleur dominante. Exemple : Histogramme de probabilité RGB.

- Histogramme de texture qui s'intéresse aux motifs de couleurs.

Exemple histogramme de Hough.

- Histogramme de forme qui s'intéresse à la forme des objets.
  Exemple : Histogramme de Fourier.

1.3.2 Méthodes de description locale par points d'intérêts

La méthode habituelle consiste à sélectionner certaines primitives afin d'effectuer une analyse locale sur eux. Il faut absolument détecter un nombre suffisant de points d'intérêts. Ces points doivent être bien localisés, distinctifs et stables par rapport à certaines transformations. Plusieurs travaux ont été établis dans la littérature et ont mené à une application fiable et robuste des détecteurs et descripteurs [Awad, 2016].
Exemples : Harris, SIFT, SURF, BRISK,...

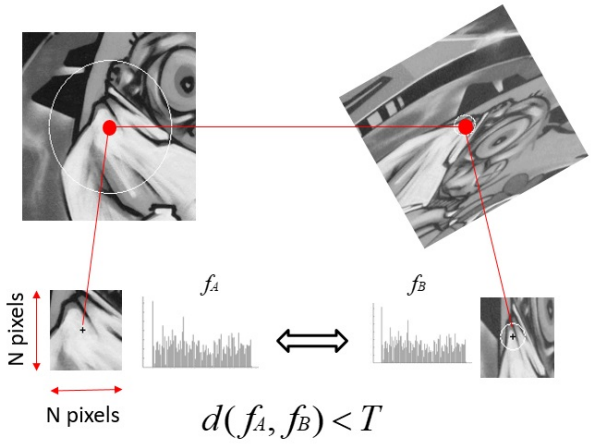

Figure 1.6. Illustration de la reconnaissance en utilisant les caractéristiques

Le but de présenter des caractéristiques locales invariantes est de fournir des représentations qui facilitent la mise en correspondance, de manière efficace, les structures locales entre des images subissant des transformations telles que les images qui composent les vidéos. Pour cette raison, que dans ce travail de thèse on s'est focalisé sur les

travaux qui ont étudié le processus d'extraction des caractéristiques complet avec ses différentes étapes : étape de détection des points d'intérêts, description des points d'intérêts et mise en correspondance de ces points.

Comme illustre la figure 1.7, les différentes étapes à réaliser dans le processus d'extraction des caractéristiques sont : la localisation des points d'intérêts, considérer la région autour des points d'intérêts afin de calculer le descripteur puis la mise en correspondance.

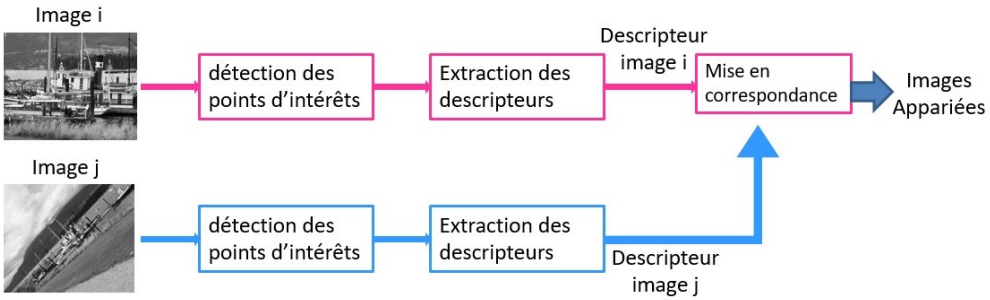

Figure 1.7. Architecture générale du processus de description locale par points d'intérêts

1.3.2.1 Détection des points d'intérêts

La détection de points d'intérêts est est une étape primordiale dans le processus d'extraction des caractéristiques locales. En effet, elle permet de localiser les points les plus importants dans chacune des images de la vidéo tout en prenant en considération le changement que peuvent subir les images appartenant à une même scène. Plusieurs méthodes dans la littérature ont été réalisées pour effectuer la détection. Ainsi, nous allons dans les paragraphes qui suivent, monter l'évolution de certains détecteurs qui

sont considérés référence dans la littérature.

a - Détecteurs de coins

Dans ce paragraphe, nous allons présenter quelques détecteurs qui ont bien contribué dans la littérature et qui sont appelés : les détecteurs de coins. Le premier détecteur est celui de Harris [Harris, 1988]. Ce détecteur est resté une référence jusqu'à nos jours bien qu'il est l'un des premiers proposés dans la littérature. Il est basé sur le calcul de gradient à l'intermédiaire des matrices d'auto-corrélation. Son principe de détection est illustré dans la figure 1.8. Il est caractérisé par sa robustesse face à la rotation, à la translation et aux changements de luminosité. Il présente un taux de répétabilité élevé (c'est-à-dire le fait que les mêmes points d'intérêts sont détectés sur deux images différentes numériquement mais représentant la même scène) bien qu'il est sensible au bruit et au changement d'angle de vue.

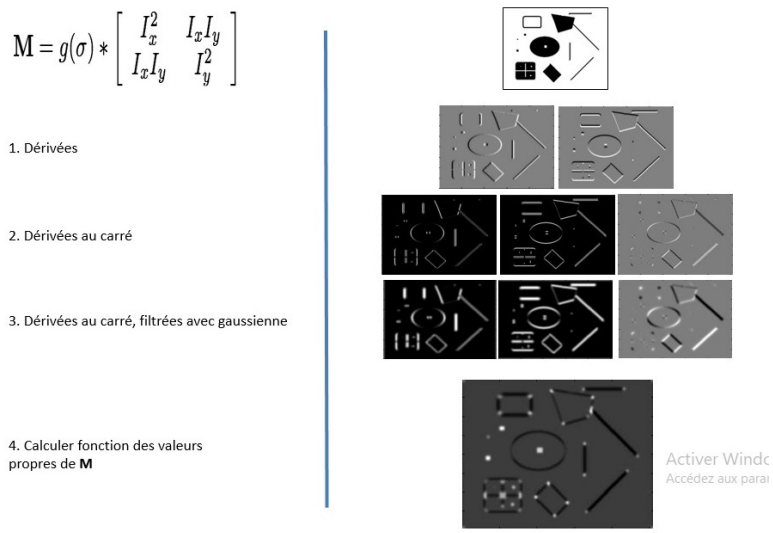

Figure 1.8. Illustration du principe de détecteur des coins "Harris". Source : [Tuytelaars et Mikolajczyk, 2008].

Le détecteur SUSAN [Smith, 1997] a été introduit dans le but de corriger cet in-

convénient. Mais, il n'a pas réussi à obtenir autant de répétabilité et de robustesse. Un détecteur très similaire a été proposé dans [Rosten, 2006], FAST est basé sur une comparaison entre le pixel central et les pixels du voisinage proche en termes d'intensité. Le résultat de comparaison sera sous forme de fraction entre nombre de pixels similaires et le nombre de pixels considérés. D'après sa conception, il était destiné aux applications temps réel avec une faible perte au niveau de précision en comparaison avec SUSAN.

b - SIFT : "Scale Invariant Feature Transform"

La première étape du détecteur SIFT [Lowe, 2004] consiste à la détection de l'extrema de l'espace d'échelle. En effet, c'est une étape de préparation, qui consiste à identifier l'espace d'échelle qui présente une invariance par rapport au redimensionnement de l'image. C'est un ensemble d'images lissées et ré-échantillonnées. Le lissage se fait utilisant un masque gaussien :

$$G(x, y, \sigma) = \frac{1}{2\Pi\sigma^2} e^{-\frac{(u^2+v^2)}{2\sigma^2}} \tag{1.5}$$

La base d'espace d'échelle L est définie donc par une convolution entre la gaussienne d'échelle $G(x, y, \sigma)$ et l'image d'entrée $I(I, y, \sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(I, y, \sigma) \tag{1.6}$$

Les emplacements des points clés sont les maxima et les minima du résultat de la différence de gaussiennes (DoG) appliquée dans l'échelle spatiale sur l'espace d'échelle. Pour la différence de gaussienne, à chaque échelle σ on calcule :

$$D(x, y, \sigma) = G(x, y, k\sigma) - G(x, y, \sigma) \tag{1.7}$$

Dans le but de déterminer le maximum et le minimum local, chaque point d'échantillonnage est ainsi comparé avec ses huit voisins dans l'image courante et aux neuf homologues dans l'échelle en dessus et en dessous. Ce point n'est alors choisi comme extremum

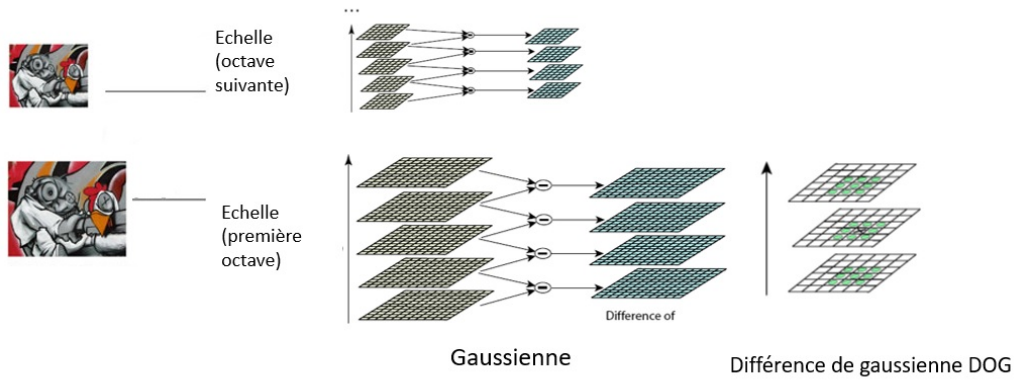

Figure 1.9. Extraction d'extrema locaux dans la fonction DOG

que lorsqu'il est plus grand ou plus petit que tous ses voisins.

Après avoir détecté les points clés candidats, l'étape suivante consiste à faire une interpolation détaillée des données pour savoir la position, l'échelle et le ratio de la courbure principale. Cette information permet de rejeter quelques points ayant un faible contraste ou qui ne sont pas bien localisés sur le contour. L'implémentation initiale consiste à localiser simplement les points clés sur la position et l'échelle du point échantillon central. Mais une méthode est utilisée pour interpoler une fonction quadratique 3D aux points échantillons locaux afin de déterminer la position interpolée du maximum. Ces expériences ont montré une amélioration d'appariement et de stabilité.

La méthode de Brown et Lowe [Lowe, 2004] utilise le développement de Taylor (jusqu'au terme quadratique) de la fonction d'espace d'échelle D(x ; y ; 6).

$$D(x) = D + \left(\frac{dD}{dX}\right)^T X + \left(\frac{1}{2}\right) X^T \frac{d^2}{dX^2} X \tag{1.8}$$

D et ses dérivées sont évaluées au point échantillon et X = (x ; y ; 6) est l'offset de ce point. La position de l'extremum est déterminée en prenant les dérivées de la

fonction selon X.

$$\hat{x} = -\left(\frac{d^2D}{dX^2}\right)^{-1}\left(\frac{dD}{dX}\right) \tag{1.9}$$

Puis, ils procèdent à éliminer les points de faibles contrastes et ceux situés sur les arêtes. Pour Lowe, les points de contraste faible sont ceux qui vérifient :

$$D(x; y; 6) \prec 0.03 \tag{1.10}$$

Les points situés sur les arêtes seront éliminés car la fonction DOG y prend des valeurs élevées, ce qui peut donner naissance à des extrema instables très sensibles au bruit. Un pic dans la fonction DOG aura une large courbure principale au long du contour mais une petite dans la direction perpendiculaire. La courbure est représentée par les valeurs propres de la matrice Hessienne 2*2 calculé à l'emplacement et l'échelle du point clé.

$$H\begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix} \tag{1.11}$$

Les expérimentations de [Lowe, 2004] prennent la valeur r=10 pour vérifier que :

$$\frac{Tr(H)^2}{Det(H)} \prec \frac{(r+1)^2}{r} \tag{1.12}$$

Si cette condition est vérifiée alors le point sera considéré comme point d'intérêt sinon il sera rejeté.

c - SURF " Speeded Up Robust Features"

La méthode d'extraction de points d'intérêts SURF proposée par Bay et al, [Bay, 2008] présente un détecteur de points d'intérêts appelé Fast-Hessien. Ce dernier se base sur une approximation du filtrage gaussien qui se charge essentiellement de minimiser le temps de calcul. Donc, pour un point x et une échelle on obtient :

$$H_\sigma(x) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \tag{1.13}$$

Dans cette méthode, les dérivées secondes des fonctions gaussiennes sont approximées par les simples filtres illustrés dans la figure 10.

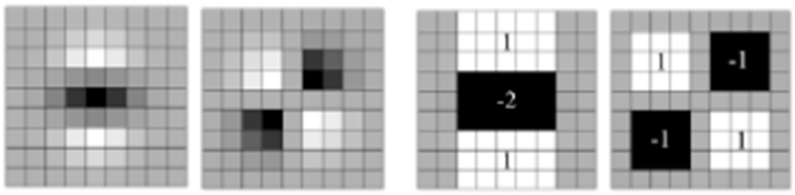

Figure 1.10. Illustration des dérivées secondes de gaussiennes et des filtres d'approximations utilisés par le détecteur SURF

Les convolutions obtenues utilisant des dérivées régularisées (par un filtrage gaussien) sont notées Dxx, Dxy et Dyy. Afin de garder la cohérence dans le filtrage, Bay et al. Utilisent un filtre approximé initial de taille $9 * 9$ correspondant à un filtre gaussien d'écart type = 1, 2. Cette cohérence est assurée par l'équation 1.14 :

$$det(H_{approx}) = D_{xx} - D_{yy} - (0.9D_{xy})^2 \tag{1.14}$$

d - ASIFT "Affine Scale Invariant Feature Transform "

Ce détecteur est considéré comme une extension de détecteur SIFT. Il a été proposé par [Morel, 2009] dans le but d'améliorer les performances du détecteur SIFT. Son principe est basé essentiellement sur la méthodologie efficace du détecteur SIFT. Sauf qu'il étend la simulation de SIFT pour une meilleure efficacité aux transformations affines et plus précisément au changement d'angle de vue.

En effet, chaque prise de vue est simulé pour des angles échantillonnés puis enregistrés

sous forme de sous forme du descripteur SIFT. Son détecteur, étant basé sur la phase de détection du SIFT, prenant avantage de ses qualités. Il améliore la robustesse surtout pour les angles de vue allant jusqu'à 80° . Ce qui favorise notamment la répétabilité. La figure 1.11 illustre un aperçu sur la robustesse de ASIFT face aux transformations affines.

Comme récapitulation, on peut noter que ce détecteur garantit la représentation des scènes, de grande taille, d'une façon simple et efficace tout en maintenant un taux d'appariement correct. Son inconvénient majeur est qu'il est trop long et par la suite ne peut pas être utilisé en temps réel ou par les applications limitées par la contrainte du temps.

Figure 1.11. Aperçu de la robustesse d'ASIFT face aux transformations affines Source : [Yu et Morel, 2009].

1.3.2.2 Description des points d'intérêts

Bien que les points d'intérêt détectés ont un contenu considéré très informatif. Plusieurs travaux dans la littérature considèrent que ces informations obtenues sont insuffisantes pour la mise en correspondance et que le passage au traitement des informations relatives aux voisinages locaux autour des différents points détectés. Ainsi,

le résultat obtenu aura les caractéristiques nécessaires pour une mise en correspondance efficace. Les caractéristiques les plus importants sont : la rapidité de calcul, l'invariance aux différentes transformations (rotation, angle de vue, échelle,..), le bruit et les occultations.

Plusieurs variantes des caractérisations locales ont été proposées dans la littérature [Dong, 2013][Krig, 2016]. La caractérisation en se basant sur les moments (moments de Hu et moment de Zernike) a été proposée initialement vu que les moments ont l'avantage de garder une invariance face aux changements de rotation, translation et changements d'échelle isotrope. Tans dis que la qualité des résultats se dégrade fortement lors d'un changement d'angle de vue ou changements d'échelle anisotrope. Il y a aussi une caractérisation utilisant le domaine fréquentiel tel que les ondelettes ou les transformées de Fourier. Il y a encore des variantes basées sur les histogrammes. Celle-ci peuvent être groupées en deux classes : celles basées sur les histogrammes d'intensité et celles basées sur les histogrammes de gradients orientés appelées HOG. Pour les descripteurs basés sur les histogrammes d'intensité, ils ont l'avantage d'une construction rapide et peu couteuse. Mais, leur utilisation n'a pas montré une efficacité en comparaison avec les méthodes utilisant les histogrammes de gradients orientés. Parmi ces derniers on va citer quelques-uns qui sont considérés comme références dans la littérature :

a - Descripteur SIFT

Après la première étape de détection des points d'intérêt et qui a été détaillée dans le deuxième chapitre, vient l'étape d'extraction des descripteurs. Puis, vient la mise en correspondance entre les images en se basant sur les points détectés ainsi que leurs descripteurs. Pour l'extraction des descripteurs, l'idée était de commencer par la collection des directions et magnitudes des gradients puis l'extraction des orientations les plus dominantes. La magnitude et l'orientation des gradients sont calculées à l'aide

des formules suivantes :

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad (1.15)$$

$$\theta(x,y) = tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \quad (1.16)$$

Ensuite, un histogramme sera créé, cet histogramme est composé de 360 degrés d'orientations qui sont devisés en 36 (regroupements de 10). Les pics dans cet histogramme correspondent aux orientations dominantes. Des nouveaux points d'intérêts supplémentaires sont créés pour les directions correspondantes aux maximums dans l'histogramme ainsi qu'aux directions dont leurs valeurs dépassent 80% de la valeur maximale. Les nouveaux points d'intérêts ont la même localisation et échelle de même que l'orientation principale mais diffèrent dans la direction. Autour de ce point, on commence par modifier le système de coordonnées local pour garantir l'invariance à la rotation, en utilisant une rotation d'angle égal à l'orientation du point-clé, mais de sens opposé. Ensuite 16 histogrammes locaux, représentant l'orientation locale du gradient sur des zones de 4*4 pixels autour du point central, sont établis. Chaque histogramme contient 8 bins qui représentent les 8 orientations principales entre 0 et 360 degrés. Par la suite, les histogrammes obtenus sont normalisés afin d'assurer une invariance aux changements d'illumination. On obtient finalement des descripteurs SIFT ayant une dimension égale 4*4 x 8 = 128 bins.

b - Descripteur SURF "Speeded up robust features"

A la différence du SIFT qui utilise les HOG pour décrire les points d'intérêt, le SURF se base sur le calcul des sommes de réponse d'ondelette de Haar. Les réponses sont représentées par des points dans l'espace. L'orientation locale est calculée en sommant les réponses verticales et horizontales incluses dans une zone de taille $\pi/3$, comme le montre la figure 1.13 :

Afin d'obtenir une invariance à la rotation et à l'échelle, l'algorithme reprend les

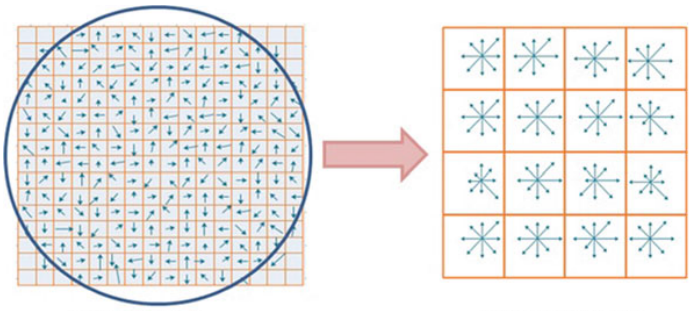

Figure 1.12. Illustation de la création d'histogramme de gradiants Gradiants d'images (à gauche) - Descriptrur de point d'interêt (à droite) [Awad, 2016]

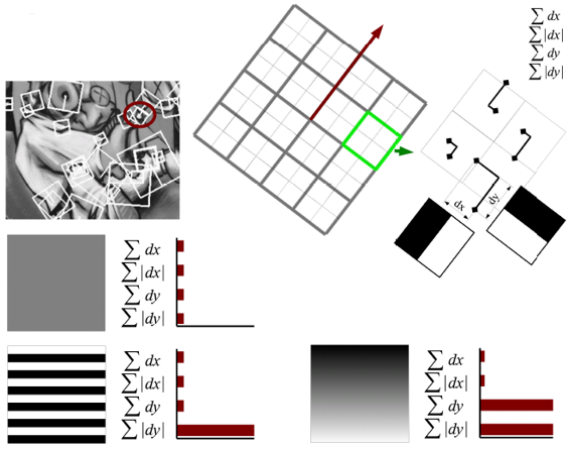

Figure 1.13. Illustration du descripteur de points d'intérêts SURF [Awad, 2016]

mêmes techniques que SIFT. Après la détermination de la valeur d'échelle et de l'orientation principale du point d'intérêt, une région d'intérêt est découpée en bloc de 4*4. Dans chaque bloc, des descripteurs simples sont calculés formant un vecteur v défini par :

$$v = (\sum dx, \sum dy, \sum \mid dx \mid, \sum \mid dy \mid) \tag{1.17}$$

Avec (et respectivement) sont les réponses d'une analyse par ondelettes de Haar dans la direction horizontale (et respectivement verticale). Cela conduit à l'obtention d'un vecteur descripteur ayant une dimension de 4*4*4=64 (SURF-64). De même, il est possible de construire un descripteur de 128 éléments (SURF-128) en calculant les termes suivants de façon séparée.

$$\sum d_x et \sum \mid d_x \mid pour\, d_x \prec 0\, et\, d_x \geq 0 \tag{1.18}$$

$$\sum d_y et \sum \mid d_y \mid pour\, d_y \prec 0\, et\, d_y \geq 0 \tag{1.19}$$

Ensuite, vient l'étape de mise en correspondance qui se base sur une minimisation de la distance euclidienne inter-descripteur.

c - GLOH "Gradient Location Orientation Histogram"

Ce descripteur a été proposé par Mikolajczyk et al., [Mikolajczyk, 2005] puis amélioré par Chandrasekhar et al [Chandrasekhar, 2009] . L'idée principale de ce descripteur est de construire un histogramme de gradients orientés qui sera représenté au sein d'un plan circulaire. Le GLOG est construit ainsi, en 17 zones d'analyse, 3 paramètres radiaux (r1, r2 et r3) et 8 paramètres angulaires.

La figure 1.14 montre un exemple illustratif du masque d'analyse de descripteur GLOH Ce masque est composé des trois cercles (avec des rayons r1,r2 et r3). Les

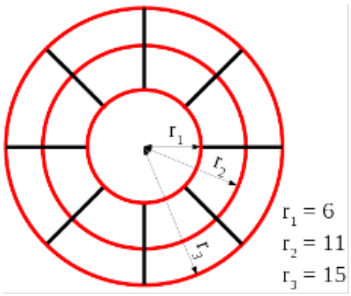

Figure 1.14. Illustration du masque d'analyse du descripteur GLOH

deux cercles les plus grandes sont divisées en huit zones (chacune de $\pi/4$). Enfin pour chacune de ces zones, un histogramme de gradients orientés est construit suivant 16 classes (suivant des intervalles de $\pi/8$). Et l'histogramme résultat sera composé de 272 éléments. Ces données sont seuillées et normalisées. Après plusieurs études, les études comparatives ont prouvé que ce descripteur améliore dans certains cas le descripteur traditionnel SIFT.

d - DAISY

Poposé par [Tola et al., 2009], le descripteur DAISY a été inspiré à partir des avantages de deux descripteurs performants : SIFT [Lowe, 2004] et GLOH [Mikolajczyk, 2005]. En effet, il a été construit en se basant sur un modèle permettant d'obtenir une invariance à l'échelle en se basant sur une analyse multi-échelle autour du voisinage local du point d'intérêt détecté. La figure 1.15 montre une illustration du principe de construction de ce descripteur.

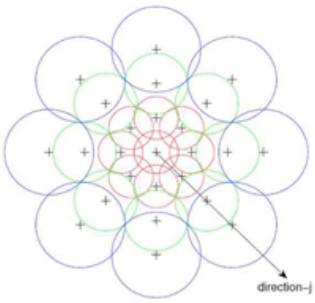

Figure 1.15. Illustration du principe de construction du descripteur DAISY [Tola, 2009]

Chaque cercle dans cette figure représente une région. Chaque rayon est proportionnel à l'écart-type du noyau gaussien. Le signe + indique les emplacements ou les centres des cartes convoluées d'orientations sont échantillonnés. Les descripteurs sont ainsi calculés dans ces emplacements. Une superposition entre les différentes

régions montre une certaine transition entre les régions et la robustesse à la rotation. Ainsi les rayons des régions extérieures qui sont plus grands permettent d'avoir un échantillonnage égal en relation avec l'axe de rotation. Ce qui confirme de plus la robustesse à la rotation.
Les expérimentations établies par [Tola, 2009], montrent que ce détecteur présente une invariance à la translation ainsi qu'à la rotation et les transformations géométriques. Contrairement aux différents descripteurs cités précédemment, DAISY présente uniquement une étape de description autour des points. Il n'a pas présenté une étape de détection des points d'intérêts. Il a été souvent nommé par le descripteur dense vu qu'il pourra être calculé sur tous les pixels de l'image. Mais dans la majorité des cas il a été utilisé dans le contexte de points d'intérêts couplé avec le détecteur Harris.

1.3.2.3 Mise en correspondance des points d'intérêts

La phase de mise en correspondance est la dernière étape dans le processus d'extraction des caractéristiques. En effet, les deux étapes précédentes sont considérées comme pré-traitement pour faciliter cette tâche. Les trois phases sont généralement dépendantes les unes des autres, car la méthode de mise en correspondance prend généralement en considération le type de détecteur et les différentes caractéristiques, surtout celles des types d'invariance. En effet, de nombreuses méthodes de mise en correspondance de points d'intérêts existent dans la littérature. Dans cette partie, nous allons présenter les principales catégories existantes. Vu que notre travail est basé essentiellement sur la description locale autour des points d'intérêts, nous allons étudier uniquement les méthodes de mise en correspondance basée sur ce type de caractéristique. Les principales catégories de mise en correspondance sont basées essentiellement sur des techniques de:

- corrélation
- relaxation

- multi-résolution (hiérarchique)
- ajout de contraintes spatiales

Ainsi, trois cas d'appariements sont alors possibles :
- Les bons appariements (appelés également inliers) qui déterminent la qualité et la précision de mise en correspondance d'une méthode.
- Les faux appariements (aussi appelés outliers), ce sont les mauvais appariement, qui détériorent les performances des applications "haut niveau". L'objectif est donc d'en diminuer le nombre de ces faux appariements.
- Les points qui ne s'apparient pas, sont généralement issus d'un processus cherchant à diminuer les outliers, ils ont l'avantage de ne pas pénaliser certaines applications surtout celles de "haut niveau".

a - Appariement par corrélation

Les méthodes basées sur le principe de corrélation sont principalement utilisées dans l'analyse de l'information des intensités pour la mise en correspondance. Leur principe consiste à déterminer, pour le voisinage d'un point x_1 d'une première image, la corrélation maximale (distance minimale) avec le voisinage issu de la seconde image. Ce calcul permet donc d'extraire le point x_2 formant ainsi le couple (x_1, x_2) présentant la meilleure ressemblance au sens de la corrélation. Ce schéma de la figure 1.16 résume une telle mise en correspondance.
Afin d'optimiser cette méthode, une estimation de la position de x_2 peut être introduite. Nous déterminons alors les mesures de corrélation à l'intérieur d'une zone de recherche, et non plus sur l'image entière.

b - Appariement par relaxation

Proposée par Hummel et Zucker [Hummel] en 1983, puis améliorée par Sidibe et al. [Sidibe, 2007] en 2007, la mise en correspondance par relaxation se base sur une

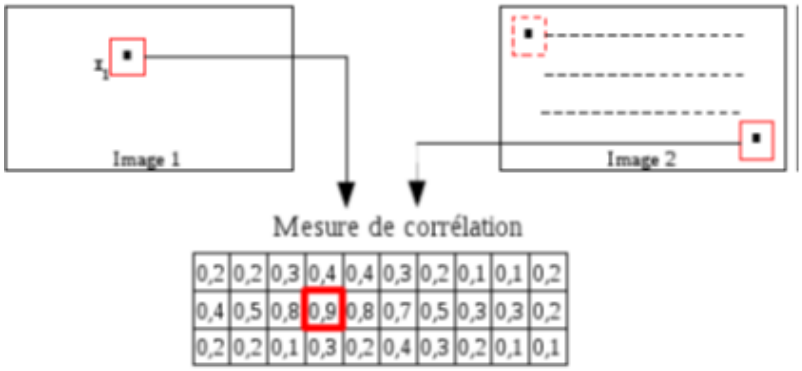

Figure 1.16. Principe de la corrélation Recherche du point issu de la ressemblance [Brochier, 2011].

fonction de probabilité d'appariement. Le principe est de calculer la probabilité qu'un point x_i soit apparié avec un point x_j connaissant les appariements de ses voisins. Cette probabilité, notée $p_i(j)$, est tout d'abord initialisée, puis est mise à jour de façon itérative jusqu'à l'obtention d'un point stationnaire $p_i^k(j)$. La mise à jour se base sur une fonction de compatibilité q_i, définie dans le voisinage V_i du point x_i. Il existe différents modèles d'appariement par relaxation, celui préconisé par Hummel et Zucker est défini par :

$$p_i^{k+1}(j) = \frac{p_i^k(j)q_i^k(j)}{\sum p_i^k(j)q_i^k(j)} \, avec \, q_i^k(j) = \sum_g W_{ig} \sum_h p_{ig}(j,h) p_g^k(h) \qquad (1.20)$$

ou $p_{ig}(j,h)$ est la probabilité que le point x_i soit apparié avec x_j sachant que le point x_g est apparié avec x_h. Le coefficient w_{ig} permet de quantifier l'influence de x_g sur x_i.

c - Appariement par multi-résolution

En s'appuyant sur une mesure de corrélation de type SSD (la somme des distances au carrée), Chen et Hung proposent en 2007 [Chen, 2007] une méthode multi-résolution de mise en correspondance. Cette dernière repose sur la construction d'une pyramide

constituée d'images successives. L'image initiale I_l caractérise la base de la pyramide. Les étages supérieurs, $I_{l-1} \dots I_1$, sont calculés par lissage et échantillonnage, pour conclure par l'image I_0 représentant le voisinage du point d'intérêt. La valeur d'intensité du point $(i; j)$, résultant de l'échantillonnage de I_1 vers I_{l-1} est déterminée par :

$$I_{l-1}[\frac{i}{2}, \frac{j}{2}] = \frac{1}{4}(I_l[i, j], +I[i+1, j] + I_l[i, j+1] + I_l[i+1, j+1]) \tag{1.21}$$

Le schéma de la figure 1.17 représente la pyramide une fois construite, pour une image initiale de taille NxN.

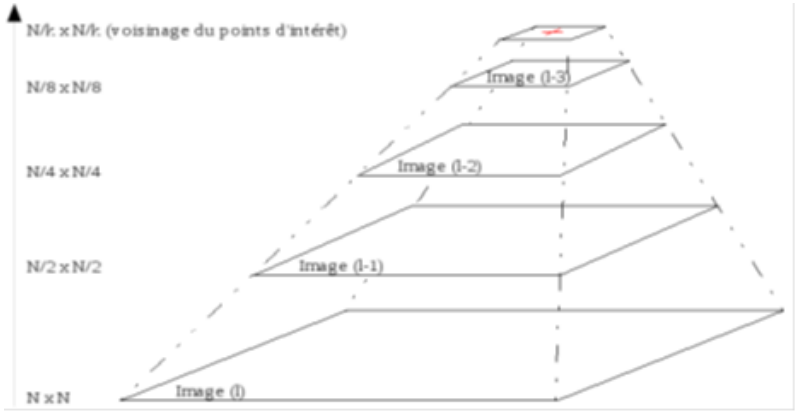

Figure 1.17. Exemple de pyramide constituée d'images successives construites par échantillonnage. Leurs tailles respectives sont mentionnées sur la gauche [Brochier, 2011].

Les mesures de corrélation se calculent de façon hiérarchique. Le processus débute par l'image située au sommet de la pyramide (image I_0) et se termine par l'image haute résolution (image de départ I_l). Chaque transition $I_{l-(k+1)} \rightarrow I_{l-k}$ entraine une augmentation de la taille de la fenêtre d'analyse V du voisinage du point ainsi que celle de la zone de recherche S de la SSD.

d - Appariements avec ajout des contraintes spatiales

Plusieurs travaux ont considéré que l'ajout des contraintes spatiales diminue le taux d'apparition des faux appariements et nous donne un résultat plus consistant. Pour résoudre ce problème de contraintes spatiales dans l'appariement, plusieurs travaux ont été introduits. Le début, c'était l'apparition des techniques d'appariement des graphes. Plusieurs travaux ont été présentés dans ce contexte tel que [Tu, 1999] et [Torresani, 2008] qui ont formulé le problème utilisant un modèle graphique. Dans [Leordeanu, 2007] aussi une technique spectrale basée sur les graphes et la similarité entre les paires de points a apparu. Kingsbury a proposé un autre travail [Kingsbury, 2010] étroitement lié à [Leordeanu, 2007] en 2010 dans lequel on calcule une mesure de similarité basée sur les contraintes spatiales entre les paires de points d'intérêts et la mise en correspondance sera par la suite en cherchant les paires qui satisfirent l'espace de similarité. Ses différentes étapes peuvent se résumer comme suit :

- Former des groupes de points d'intérêts (fenêtres adjacentes qui ont 75% zone de Chevauchement entre eux).
- Si on considère deux images X et Y, N et M groupes de points d'intérêts sont formés dans X et Y respectivement et l'appariement entre les éléments de (Gxn) et (Gym) sera en utilisant le seuil de la distance euclidienne (SIFT), le calcul des mesures de similarités sera pour toutes les combinaisons des paires sur les points de (Gxn) qui correspondent à des points de (Gxm) puis considérer les des contraintes spatiales entre les paires de points d'intérêts [Kingsbury, 2010].

1.3.3 Discussion

La segmentation automatique en blocs des images donne des informations locales qui sont assez souvent pas assez précises pour l'étape de description. Ceci est du aux différentes lacunes de ces algorithmes de segmentation et à la nature du contenu de ces images . Plusieurs travaux dans la littérature ont utilisé les points d'intérêts pour la

description locale et leur utilisation a montré une grande robustesse face à différentes transformations que peut subir une image. Ainsi, le défi consiste à choisir le détecteur (ou descripteur) adéquat aux différents besoins vu que la conception de chacun a été dédiée à un type bien défini d'application. Dans les paragraphes précédents, nous avons présenté les différents détecteurs, descripteurs et méthodes de mises en correspondance les plus cités dans la littérature. Chacun d'eux présente des avantages et des inconvénients. Malgré l'apparition de plusieurs, SIFT [Lowe, 2004] reste une référence dans la littérature [Awad, 2016]. En effet, sa représentation est remarquable à plusieurs égards : il est soigneusement conçu pour éviter les changement de bords, orientations et échelles. Cependant, il n'est pas explicitement invariant aux transformations affines ainsi que la construction de son vecteur descripteur (vecteur de 128 éléments décrivant un patch de pixel) entraine un problème de haute dimensionnalité qui affecte un temps de calcul significativement lent. Dans le but d'améliorer ce temps de calcul lent de SIFT, les auteurs de [Bay, 2006] ont utilisé un descripteur de taille 64 au lieu de 128. Cependant, les résultats de SIFT restent incomparable dans plusieurs transformations tel que la translation, la rotation, changement d'échelle et illumination [Pang, 2012]. De même, ASIFT [Morel, 2009] est une variante qui a résolu l'invariance affine du détecteur SIFT mais il est aussi trop long et par la suite ne peut pas être utilisé en temps réel ou par les applications limitées par la contrainte du temps. Quant au descripteur GLOH [Chandrasekhar, 2009], il est aussi assez similaire a SIFT tout en étant plus performant surtout en terme de changement d'illumination mais il reste assez couteux (même plus que SIFT). Ainsi, en se basant sur ces critères, nous allons dans le prochain chapitre justifier le choix du détecteur ainsi que du descripteur de points d'intérêts que nous allons utiliser.

Conclusion

Dans ce chapitre nous avons commencé par présenter les concepts généraux des systèmes de recherche de contenu, citer les différents systèmes existants en se focalisant sur les systèmes de recherche de vidéos par le contenu. Par la suite, nous avons présenté une synthèse générale dans laquelle on a présenté les méthodes de construction de résumé de vidéos existantes. Ces méthodes peuvent être classées selon différentes catégories. Le principe général, les avantages et les inconvénients ainsi que quelques exemples de chacune de ces catégories ont été présentés. Nous avons constaté par la suite que malgré les nombreuses méthodes de construction de vidéos existantes, il y en a beaucoup qui sont basées sur la description de caractéristiques globales tandis que très peu qui ont tiré profit de la caractérisation locale malgré qu'elle pourrait être une alternative très fructueuse tout en sachant que l'étape d'extraction de caractéristique est l'étape clé dans n'importe quelle application de recherche par le contenu y compris la construction de résumé vidéos. Dans ce contexte, nous avons présenté les différentes catégories d'extraction des caractéristiques tout en se focalisant sur les descripteurs locaux, leurs avantages, leurs processus général et les différentes étapes.

Chapitre 2

CONSTRUCTION DES RÉSUMÉS DE VIDÉOS PAR MISE EN CORRESPONDANCE DES POINTS D'INTÉRÊTS ET CLASSIFICATION DES VALEURS DE RÉPÉTABILITÉ

Introduction

L'objectif principal de la construction de résumé vidéo est de faciliter la recherche de vidéos par le contenu. En effet, dans un archive vidéo, l'utilisateur pourra introduire une image requête et il pourra, par la suite, récupérer toutes les vidéos ayant un contenu similaire.

Notre objectif est ainsi, de construire un résumé qui répond aux défis suivants : être fidèle à la vidéo d'origine et contenant les objets les plus représentatifs de cette vidéo tout en minimisant la redondance.

Pour ce faire, les images de vidéos doivent être analysées afin d'en extraire certaines primitives globales ou locales. Une des méthodes de description locale les plus courantes s'appuie sur l'utilisation des points d'intérêts qui sont des primitives locales riches en contenu. Dans le but de mettre en correspondance l'ensemble des points d'intérêts détectés d'une image de la vidéo à une autre. Une description locale du point ainsi que son voisinage est mise en œuvre. Dans ce contexte, nous avons proposé une méthode de mise en correspondance prenant avantage de la description locale des points d'intérêts et de la notion des invariants géométriques. Cette méthode sera utile pour la construction du résumé statique de vidéo, sous forme des images clés, ainsi que pour l'indexation des vidéos. Nous allons, dans la suite de ce chapitre,

présenter deux méthodes d'extraction des images clés tout en détaillant les différentes étapes de chacune.

2.1 Contexte général des méthodes proposées

Les différentes méthodes que nous allons proposer dans ce travail de thèse se situent dans le cadre de la recherche de vidéo par le contenu. La figure 2.1 montre l'architecture générale du processus de recherche que nous allons suivre. Nous nous intéressons essentiellement aux deux étapes, considérées parmi les plus importantes dans ce processus, et qui sont : l'extraction des images clés et l'extraction des caractéristiques locales. Cette dernière sera utile non seulement pour l'étape de calcul de similarité dans le système de recherche mais aussi pour l'extraction des images clés. Ainsi, nos contributions se focalisent autour de ces deux étapes.

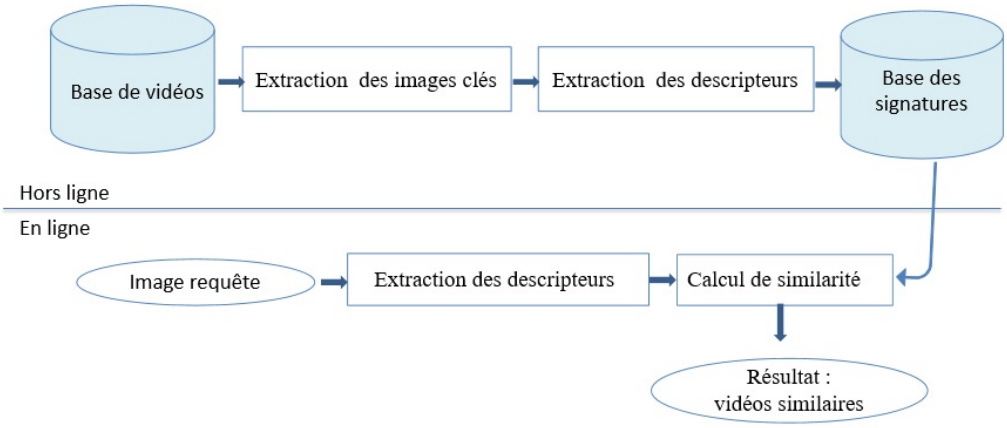

Figure 2.1. Architecture générale du système de recherche de vidéos par le contenu.

En effet, le résultat de génération du résumé de vidéo n'est pas lié uniquement à la robustesse de la méthode de construction du résumé utilisée mais aussi à la qualité de la description utilisée. L'extraction de caractéristiques est considérée comme une

étape cruciale qui influence fortement la qualité du résumé. La plupart des méthodes de sélection des images clés sont basées sur l'extraction des caractéristiques globales. Dans les méthodes que nous proposons dans ce travail de thèse, nous allons tirer profit des multiples avantages fournis par les caractéristiques locales présentées précédément. Nous présentons, dans la figure 2.2 un chronogramme du processus générique utilisé pour l'extraction des images clés basé sur la description locale.

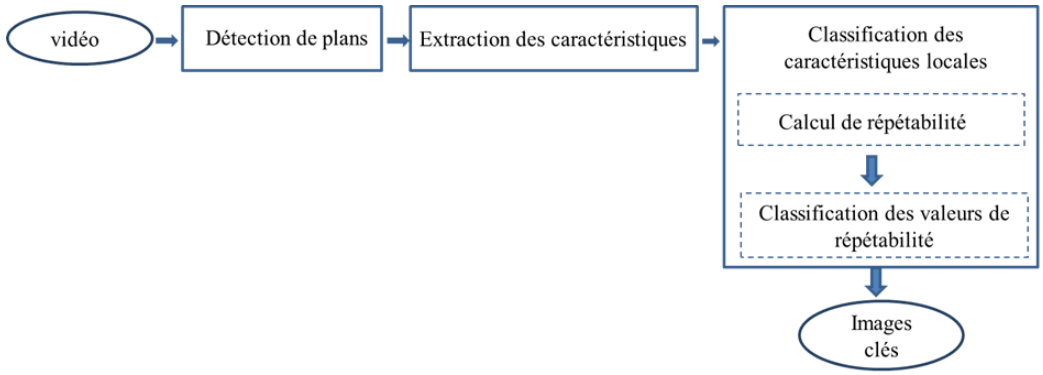

Figure 2.2. Processus général d'extraction des images clés

Nous présentons par la suite la description de chacune de ces étapes:

- Détecter les limites des plans : Dans cette phase, nous avons utilisé la méthode des histogrammes chi-2 proposée par [Cai, 2005]. La détection des bords des plans dans cet algorithme se base sur la comparaison de l'histogramme de chaque image traitée et celui de l'image de référence qui est automatiquement détectée. Nous avons opté pour cet algorithme vu sa précision de détection des bords des plans. En effet, il permet, grâce à la combinaison de l'histogramme couleur avec la distance X^2, d'éviter la sensibilité des objets en mouvements. Ceci permet de regrouper séquentiellement les images similaires.

- Extraction des caractéristiques locales : Cette étape mettra l'accent sur deux sous-étapes qui sont : celle de description locale autour des points d'intérêts

en utilisant le descripteur Local Binary Pattern LBP et celle de mise en correspondance qui est basée sur les invariants géométriques. L'étape d'extraction des caractéristiques locales sera utile non seulement pour le calcul de similarité dans le système de recherche mais aussi pour l'extraction des images clés.

- Extraction des images clés : c'est l'ensemble des images formant le résumé statique de la vidéo. Ce résumé permet une description des vidéos de la base dans le but de faciliter le processus de recherche.

2.2 Mise en correspondance par invariants géométriques (MCIG)

Dans ce paragraphe, nous allons présenter l'étape d'extraction des caractéristiques locale basée sur les points d'intérêts proposée dans cette thèse. Cette méthode est composée essentiellement de deux étapes (figure 2.3) :

- La première étape consiste à décrire l'ensemble des points d'intérêts détectés ainsi que leurs voisinages. Pour ce faire, nous avons adapté le descripteur LBP (Local Binary Pattern), qui est essentiellement un descripteur de texture [Ojala, 2000][Pietikäinen, 2011], au contexte de points d'intérêts. Dans sa forme générique, ce descripteur est connu par sa sensibilité aux rotations [Pietikäinen, 2011]. Pour cette raison, nous avons adapté sa représentation pour qu'il soit invariant à la rotation.

- La deuxième étape permet la mise en correspondance entre les points d'intérêts appartenant à deux images. Cette mise en correspondance est basée, en plus de la comparaison du voisinage local entre descripteurs, sur une comparaison basée sur les invariants géométriques.

Ces deux étapes doivent être précédées par une étape de détection des points d'intérêts. Une multitude de méthodes robustes ont été proposées pour la détection des points

d'intérêts dans la littérature. Chaque détecteur est dédié à des tâches prédéfinies. Ceci dépend essentiellement des besoins des utilisateurs. D'après la section 1.3.3, nous avons trouvé que le détecteur SIFT s'adapte mieux à notre contexte. Aussi, plusieurs études dans la littérature [Juan, 2010] [Panchal, 2013] [Karami, 2015] ont comparé SIFT à différents détecteurs. Leur comparaison a été basée essentiellement sur le critère de répétabilité qui permet l'évaluation de la stabilité des points d'intérêts détectés face à différents types de changements. Ainsi, SIFT montre sa stabilité pour tous les types de transformations et contraintes sauf dans le temps d'exécution qui est relativement lent. Cela est dû essentiellement au grand nombre de points d'intérêts qui sont appariés comparablement aux autres méthodes. Parmi ces points, un certain nombre de faux appariements existent [Juan, 2010][Li, 2015]. Ainsi, nous avons utilisé seulement l'étape de détection proposée par SIFT. Puis appliquer la méthode d'extraction des caractéristiques que nous avons proposée dans le but de minimiser, au maximum, le nombre des faux appariements. La figure 2.3 montre le processus général d'extraction des caractéristiques locales adapté.

Dans ce qui suit, nous présentons une description détaillée de chacune de ces étapes:

- Étape de description locale des points d'intérêts en adaptant le descripteur de texture LBP au contexte des points d'intérêts.
- Mise en correspondance basée sur la description locale dans un premier temps, puis sur les invariants géométriques afin d'éliminer les faux appariements générés suite à la comparaison des descripteurs locaux.

2.2.1 Description des points d'intérêts par Local Binary Pattern

L'opérateur Local Binary Pattern (LBP) est un opérateur de texture qui a été développé comme étant invariant aux variations des niveaux de gris. A la base c'est un descripteur de texture. En effet, la texture est caractérisée par une grande variation de l'intensité. D'ou, elle contient forcément un grand nombre de points d'intérêts, car par définition un point d'intérêt est un pixel de l'image qui est caractérisé par une

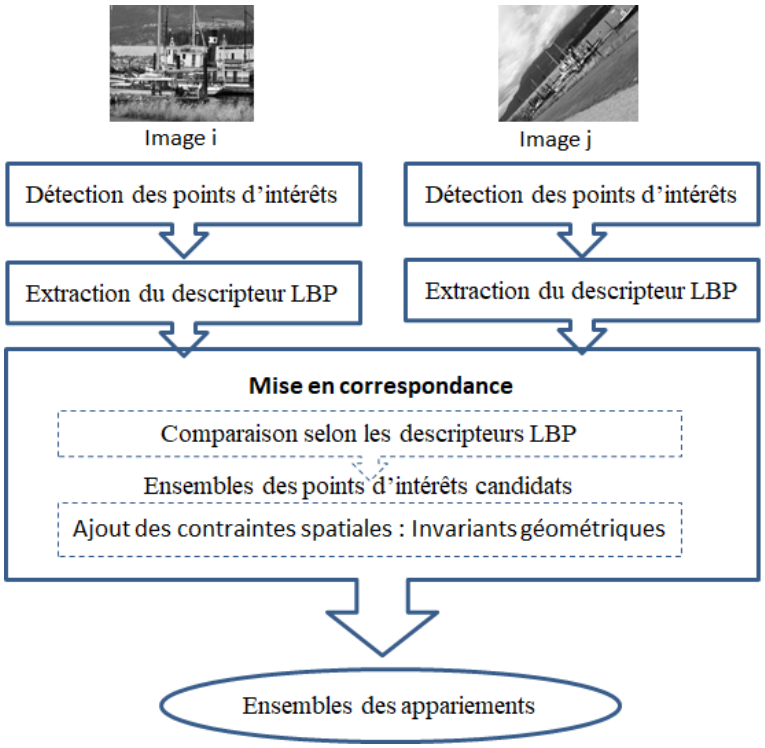

Figure 2.3. Processus général d'extraction des caractéristiques locales.

variation de l'intensité dans aux moins deux directions. Ceci, confirme que le choix de l'opérateur LBP est bien adapté pour représenter bien le contenu local autour des points d'intérêts.

L'opérateur LBP (Local Binary Pattern) est basé sur une représentation symbolique entre le pixel et son voisinage [Hafiane, 2006]. Cet opérateur a été proposé en premier lieu par [Ojala, 1996] puis développé par Harwood et al. [Harwood, 2003]. Il a montré d'excellentes performances pour la description de texture dans de nombreuses études comparatives, tant en termes de vitesse, et de capacité de discrimination. Etant indépendant de toute transformation monotones de niveaux de gris, l'opérateur est parfaitement adapté pour compléter les mesures de couleur ou à être complété par une mesure orthogonale du contraste de l'image.

Le descripteur LBP (Local Binary Pattern) dans sa forme actuelle est formé par

un ensemble de modèles locaux qui sont construits autour de chaque pixel. En effet, chaque pixel est étiqueté par le code de la texture qui correspond bien à l'échelle locale de son voisinage. Dans notre contexte, nous avons appliqué le code LBP uniquement autour des points d'intérêts détectés, et non pas pour tous les pixels appartenant aux images. Ainsi, chaque code LBP pourra être considéré comme le code qui représente au mieux le voisinage local du point d'intérêt. La distribution LBP au voisinage du point d'intérêt présente à la fois des propriétés d'ordre structurel : des primitives de textures du voisinage du point d'intérêt ainsi que des règles concernant le placement de ces primitives.

Ojala et al. [Ojala, 1996] a obtenu l'opérateur local binaire (LBP) défini dans un voisinage local d'un pixel d'une image en niveaux de gris comme la distribution conjointe des niveaux de gris de $(P + 1)$ pixels de l'image $(P > 0)$:

$$T = t(g_c, g_0, ..., g_{p-1}) \tag{2.1}$$

Où g_c correspond au niveau de gris du pixel central d'un voisinage local et g_p (p = 0,. . . .., P-1) correspond aux niveaux de gris de P pixels équidistants dans un cercle de rayon $R(R > 0)$. Les coordonnées des voisins g_p de g_c sont les suivants :

$$x_p = x_c + Rcos(2\pi p/P) \tag{2.2}$$

$$y_p = y_c + Rsin(2\pi p/P) \tag{2.3}$$

Où x_c et y_c sont les coordonnées du pixel central.

Si les valeurs des coordonnées des voisins ne correspondent pas exactement aux coordonnées des pixels, alors elles sont estimées par interpolation bilinéaire. La corrélation entre les pixels décroît avec la distance. Une grande partie de l'information sur les pixels se trouve dans leur voisinage local.

En général, cette opération est visée pour la représentation de texture, pour ceci elle est appliquée sur toute l'image. Dans notre cas, on l'a appliquée uniquement aux

points d'intérêts détectés lors de la première étape afin de produire un descripteur local particulier pour chaque point d'intérêt. Parmi les propriétés les plus importantes de l'opérateur LBP est son invariance par rapport aux changements monotones et uniformes d'éclairage.

Bien que l'opérateur LBP soit invariant par rapport aux variations des niveaux de gris, les différences sont affectées par l'échelle. Pour atteindre l'invariance par rapport à toute transformation monotone des niveaux de gris due à l'échelle, on ne considère que les signes des différences. Le code LBP sera exprimé donc comme suit :

$$code\, LBP = \{s(g_0 - g_c), ..., s(g_{p-1} - g_c)\} \tag{2.4}$$

Où

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x \leq 0 \end{cases}$$

et $g(p = 0, \ldots .., p-1)$: correspond aux niveaux de gris de P pixels voisins équidistants dans un cercle de rayon $R(R > 0)$.

La Figure 2.4 montre un exemple illustratif de calcul de l'opérateur LBP.

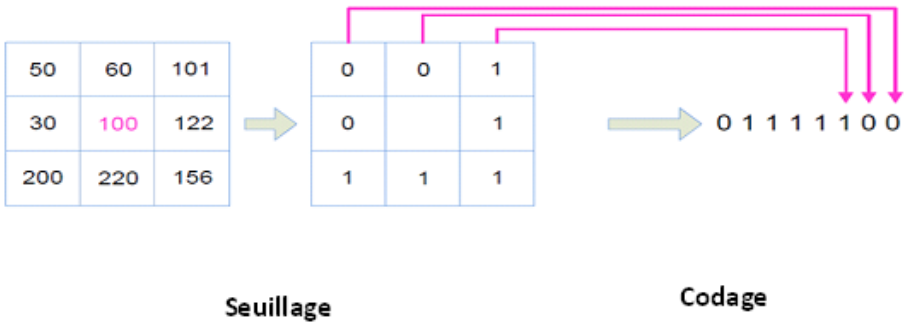

Figure 2.4. Exemple de calcul de code LBP

Malgré les différents avantages présentés, on peut noter que l'opérateur basique LBP souffre d'un majeur inconvénient qui est sa sensibilité à la rotation. Une rotation de l'image donne naturellement un code LBP différent. Plusieurs méthodes ont été

proposées dans la littérature pour permettre à l'opérateur LBP d'être invariant à la rotation. La plupart de ceux-ci sont destinés aux cas de textures. Parmi les solutions les plus populaires on peut citer le concept d'uniformité [Ojala, 2000] qui autorise un changement d'au plus deux transitions un-à-zéro ou zéro-à-un se trouvant dans le code binaire. Ce modèle est déstiné aux cas de textures régulières contenant des motifs qui se répètent. Ceci n'est pas toujours le cas pour les points d'intérêts dont leurs voisinages sont généralement caractérisés par des textures complètement aléatoires. On peut citer aussi la représentation circulaire du code LBP en cas de texture pour toute les pixels de l'image. D'après [Bahroun, 2011],lorsque une image est tournée, les pixels voisins g_p se déplacent sur le long du périmètre d'un cercle centré sur le centre g_c. En raison de cette nature circulaire du voisinage, il est devenu assez simple d'obtenir des codes LBP invariants par rapport à la rotation. Ainsi, nous nous sommes inspiré de cette alternative et nous avons considéré chaque point d'intérêt dans l'image comme un centre de rotation. Ceci semble être une convention efficace dans la détermination de l'invariance par rotation de l'opérateur LBP. D'où la conversion du code LBP, qui est sous forme d'un vecteur binaire, en forme de cercle binaire ayant comme centre le point d'intérêt considéré. Ceci rend le code LBP inchangeable quelle que soit la rotation subite. La figure 2.5 montre un exemple illustratif de la conversion du code LBP de sa forme binaire vers une représentation circulaire. Si on prend :

$$LBP = 01110100$$

Après une certaine rotation il peut se transformer par exemple en :

$$LBP = 11010001 ou en LBP = 01000111$$

Par contre, s'il est sous forme circulaire quelle que soit la rotation subite il n'y aura pas de changement. Dans le cas de texture, la description utilisant le code LBP circulaire [Bahroun, 2011] a montré des résultats performants en termes de capacité de discrimination. Dans notre cas, une petite taille de voisinage n'est pas assez

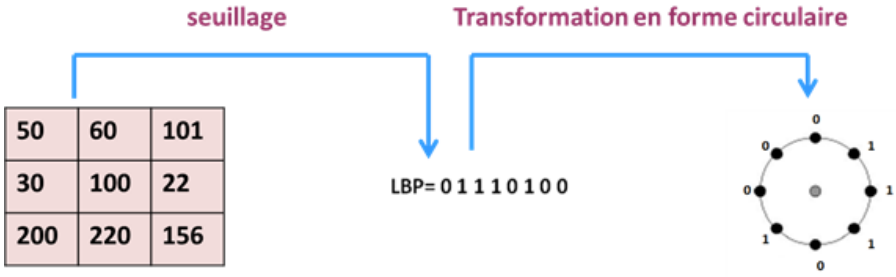

Figure 2.5. Transformation du code LBP d'un vecteur binaire à un cercle binaire

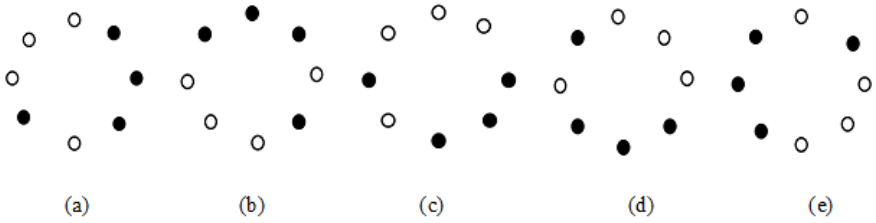

Figure 2.6. Exemples de codes LBP équivalents avec les cercles noirs et blancs correspondent à des valeurs de bits de 0 et 1 à la sortie de l'opérateur LBP.

suffisante pour capter les informations caractérisant la texture locale vu la richesse du voisinage des points d'intérêts. Dans ce contexte, on calcule le code LBP pour 5 rayons du voisinage autour du point d'intérêt. La figure 2.7 montre une illustration de calcul de code LBP pour un point d'intérêt avec $R = 1$.

Figure 2.7. Exemple de calcul de code LBP d'un point d'intérêt avec R=1

On a procédé à calculer le code LBP pour différents rayons de voisinages afin de garantir un descripteur riche en information locale autour des points d'intérêts. Nous

avons choisi de prendre 5 rayons de voisinage après plusieurs expérimentations. Ainsi, le descripteur LBP sera de taille 120 et une meilleur discrimination est obtenue.

2.2.2 Description générale de la méthode de mise en correspondance MCIG proposée

Le but dans cette partie sera de trouver pour le maximum de points détectés dans la première image leurs meilleurs correspondants dans la deuxième image. Pour ceci, nous allons suivre une méthode de mise en correspondance basée sur des filtrages des points correspondants en cascade. Afin de minimiser la complexité de notre algorithme, et aussi pour que les opérations couteuses en terme de calcul ne s'exécutent que pour un nombre limité de candidats : ceux qui réussissent à passer un test initial. La première partie sera consacrée pour ce premier test où on va détailler l'appariement initial qui est basé sur des contraintes locales. Or, ces dernières ne sont pas suffisantes vu qu'elles peuvent en résulter des faux appariements ou des ambigüités. Pour ceci, on va ajouter un deuxième test basé sur des contraintes spatiales pour minimiser les conflits et le taux d'erreurs. Ainsi, la mise en correspondance que nous avons proposée se fait en deux étapes :

- Mise en correspondance selon les descripteurs locaux
- Mise en correspondance selon des invariants géométriques

2.2.3 Mesure de similarité

Afin de mettre en correspondance deux ensembles de points d'intérêts, il est nécessaire de détecter leurs similarités en effectuant une comparaison entre les deux ensembles des points. Ainsi, les critères de sélection des appariements doivent être bien précis et surtout prennent en considération les différentes transformations que peut subir les images.

a - Comparaison des codes LBP

La comparaison basée sur les codes LBP consiste à déterminer la sélection des points d'intérêts qui se correspondent entre 2 images en ne prenant en considération que les apparences locales du voisinage autour de ces points d'intérêts. Ainsi, pour chaque descripteur d'un point d'intérêt d'une image nous cherchons le descripteur le plus similaire dans l'autre image. De cette façon, tous les points d'intérêts détectés dans la première image lors de la première phase seront comparés avec ceux de la deuxième image. Cette comparaison sera basée sur leurs descripteurs LBP représentés sous leur forme circulaire, pour chaque point la comparaison se fait pour chaque rayon et son homologue.

Le résultat de cette première étape basée uniquement sur le description locale est que pour chaque point d'intérêt de la première image, on peut avoir un ensemble de points candidats: ceux qui ont des descripteurs similaires. Dans ce qui suit, ces candidats vont subir un autre test basé sur les invariants géométriques pour améliorer le résultat obtenu étant donné qu' une caractérisation selon des paramètres locaux uniquement n'est pas suffisante pour garantir la performance de la mise en correspondance.

b - Invariants géométriques

La sélection des plus proche voisin est une étape préparatrice pour l'étape de filtrage des points candidats correspondants par invariants géométriques. Pour chacune des 2 images, nous cherchons pour chaque point d'intérêt ses voisins les plus proches. Pour ceci, nous nous sommes basés sur le calcul de la distance Euclidienne entre chaque point d'intérêt et le reste des points d'intérêts dans la même image :

La distance euclidienne est calculée comme suit :

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.5}$$

Après la sélection des plus proches voisins pour chaque point d'intérêt, on fait le tri des voisins dans un ordre croissant selon la valeur de la distance euclidienne calculée. L'ajout des contraintes spatiales, en plus de celles locales, permet d'obtenir des appariements plus robustes [Kingsbury, 2010]. En effet, les points d'intérêts ne sont considérés comme correspondants que s'ils répondent à certaines contraintes spatiales. Ces dernières exigent que les points correspondants candidats entre les deux images soient conformes en termes de structure géométrique. Pour ceci, on calcule pour chacun des points correspondants candidats et leurs voisins les plus proches des invariants géométriques qui restent valables quelle que soit la transformation subite par l'image référence.

Les transformations les plus récurrentes que peut subir l'image sont la translation, la rotation, le changement d'échelle et d'illumination. Donc, on peut considérer que le mouvement entre les deux images traitées peut être approximé par une transformation affine [Brochier, 2011]. Or, les invariants d'une géométrie affine dans le plan présentent le rapport de longueurs entre des segments colinéaires [Binford, 1993]. D'où vient l'idée de prendre en considération la relation entre chaque point d'intérêt et son voisinage le plus proche des points d'intérêts. Afin de calculer ces invariants, on a besoin pour chaque point d'intérêt ses trois voisins les plus proches.

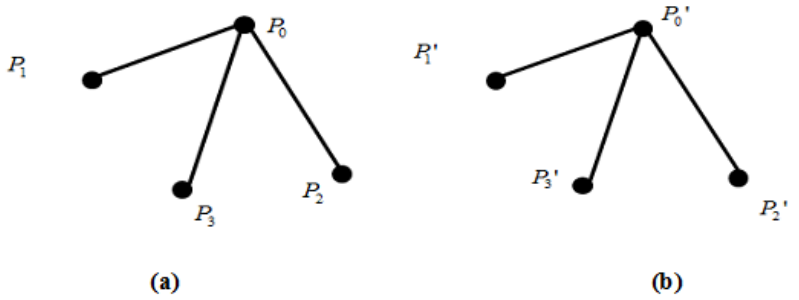

Figure 2.8. Configurations considérés pour les correspondants candidats de chacune des deux images pour le calcul des invariants géométriques.

Comme le montre les schémas de la figure 2.8, les invariants sont les coordonnées affines des deux points d'intérêts qu'on souhaite comparer P_0 et P'_0 respectivement à leurs trois voisins les plus proches (P_1, P_2, P_3) et (P'_1, P'_2, P'_3).
Ces coordonnées sont définies par le système d'équations suivant :

$$\begin{cases} a_1X_1 + a_2X_2 + a_3X_3 = X_0 \\ a_1Y_1 + a_2Y_2 + a_3Y_3 = Y_0 \\ a_1 + a_2 + a_3 = 0 \end{cases} \tag{2.6}$$

Cette configuration est mise en relief de telle sorte que les invariants sont triés selon un ordre croissant. Tel que :

$$a_1 \prec a_2 \prec a_3 \; et \; b_1 \prec b_2 \prec b_3 \tag{2.7}$$

Par la suite le processus d'appariement ne s'effectue que lors de la satisfaction d'un seuil qui a été choisi après une phase d'expérimentation sur les deux bases généralistes présentées lors du chapitre suivant:

$$\mid a_1 - b_1 \mid \leq S \tag{2.8}$$

$$\mid a_2 - b_2 \mid \leq S \tag{2.9}$$

$$\mid a_3 - b_3 \mid \leq S \tag{2.10}$$

Le choix du paramètre S à accorder aux différences entre les invariants est une étape importante qui influence le résultat d'appariement en termes de précision. De ce fait, nous allons, expérimentalement, optimiser le choix de ce paramètre en le variant dans un intervalle de 1 à 2, avec un pas de 0.05. On considère, à la fin la valeur de S qui optimise le taux de précision pour tous les types de transformations que peut subir l'image. Cette façon de choisir la valeur du seuil permet de prouver la pertinence et la fiabilité de la méthode proposée en ne permettant d'apparier que les points d'intérêts qui semblent quasi-ressemblants. Dans ce contexte, nous allons calculer

la valeur moyenne de précision (pour chaque type de transformation) en fonction de différentes valeurs du seuil. En se basant sur l'influence de la valeur du seuil, on choisit celle qui maximise le résultat.

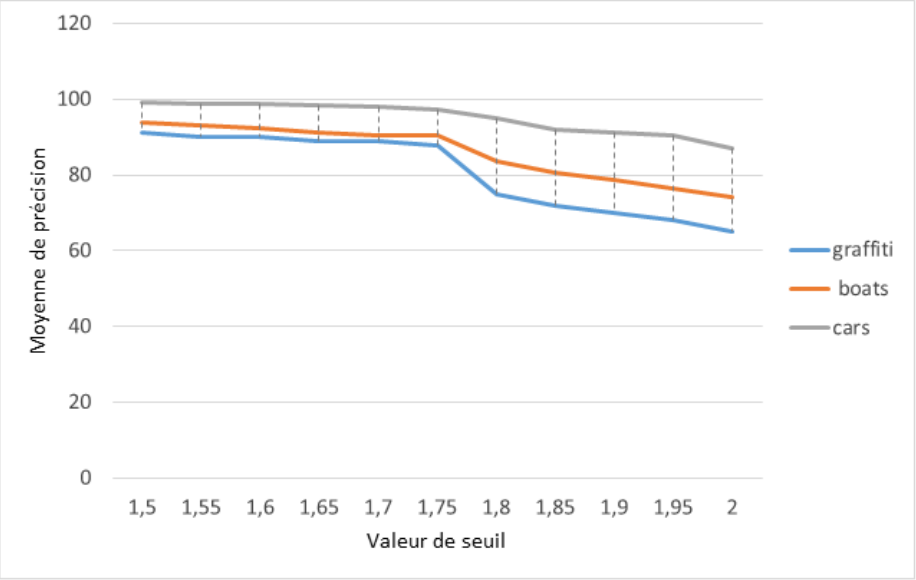

Figure 2.9. Optimisation de la valeur de précision pour la détermination de la valeur du seuil pour différents types de transformation (image "Graffiti" : changement d'angle de vue, images "boats" : couplage rotation+ changement d'échelle et "cars" : changement de luminance).

En regardant le graphique de la figure 2.9, on peut remarquer que la précision est stable jusqu'à la valeur 1,75. Au-delà de cette valeur, la moyenne de précision commence à se dégrader. De ce fait, on peut considérer que cette valeur (S=1,75) comme un bon compromis entre une stabilité de la précision et un bon nombre d'appariements. En effet, une accentuation de cette valeur cause une perte de précision et une valeur inférieure de ce seuil impose une grande sélectivité dans le processus de mise en correspondance et par la suite un nombre faible d'appariements. Sachant que la valeur moyenne de précision de SIFT et SURF respectivement : 61.6 et 75.8 pour les images "Graffiti", 84 et 82.2 pour les images "boats" et 95.6 et 96.5 pour les images "cars",

ce qui confirme un bon choix de la valeur de S.

2.2.4 Algorithme et complexité de la méthode de mise en correspondance MCIG

L'algorithme de la méthode d'appariement des points d'intérêts proposée est récapitulé dans la figure 2.3. Nous allons dans ce paragraphe, estimer la complexité de cet algorithme. Pour ce faire, nous commençons par calculer la complexité de chacune des différentes étapes de cette méthode. L'étape d'extraction des descripteurs LBP ainsi est de l'ordre de O(n) tel que n est le nombre de points d'intérêts pour chacune des images. L'étape qui permet la comparaison des descripteurs locaux, il est de l'ordre $O(n^2)$. La complexité de l'étape permettant la comparaison basée sur les invariants géométriques est estimée à $O(n^3)$. Le coût global est celui du sous algorithme ayant la complexité la plus élevée. Ainsi, la complexité de cet algorithme est évaluée à $O(n^3)$.

2.3 Extraction d'images clés basée sur la construction de la table de répétabilité (EICCTR)

Dans la suite de ce chapitre, nous allons décrire les différentes méthodes d'extraction d'images clés que nous avons proposées dans ce travail de thèse. Au début, nous avons proposé une méthode basée sur la description locale et la classification Hiérarchique ascendante. Afin d'améliorer les résultats de la classification, nous avons opté pour la réduction de la dimension en utilisant l'analyse en composante principale ACP avant de passer au processus de classification. Cette méthode a donné de bons résultats comparée à certaines méthodes existantes dans la littérature.

2.3.1 Description générale de la méthode d'extraction des images clés EICCTR proposée

La méthode d'extraction d'images clés qu'on propose initialement dans cette thèse est basée principalement sur trois étapes. Comme pré-traitement, on propose de segmenter la vidéo en plans. Pour se faire, on a utilisé la méthode proposée par Cai et al. [Cai, 2005]. Par la suite, pour chaque plan, on applique la méthode proposée décrite dans la figure 2.10. La deuxième étape consiste à extraire les descripteurs locaux pour toutes les images de chaque plan. Par la suite, on construit une table de répétabilité dans laquelle on stocke les valeurs de répétabilité inter-images appartenant au même plan. Dans ce contexte, on a tiré profit de la méthode d'appariement de points d'intérêts MCIG décrite dans la section 2.2. Comme le tableau résultant est de grande dimensionnalité. Il est de taille (N*N) avec N est le nombre d'images appartenant à chaque plan. Dans la troixième étape, on a utilisé l'analyse en composante principale ACP pour réduire la dimension des tables de répétabilité et minimiser les redondances vu le contenu similaire entre les images successives de la vidéo. Ce qui nous a facilité la dernière étape permettant classification de la table étant donné que la classification avec une basse dimension est plus efficace que dans une grande dimension. La classification hiérarchique ascendante CAH est utilisée pour regrouper les images ressemblantes en classes. Ainsi, les centres de classes seront les images clés. La figure 2.10 décrit le processus général de la méthode EICCTR proposée.

2.3.2 Construction de la table de répétabilité

Après avoir localisé les points d'intérêts pour toutes les images de chaque plan, nous allons construire une matrice, contenant les valeurs de répétabilité, qu'on a appellé 'Table de répétabilité'. En effet, la répétabilité est un critère qui permet de caractériser la stabilité des points d'intérêts détectés sous différentes variations possibles ainsi que sous l'effet du bruit. Ce sont les points détectés qui doivent être obtenus

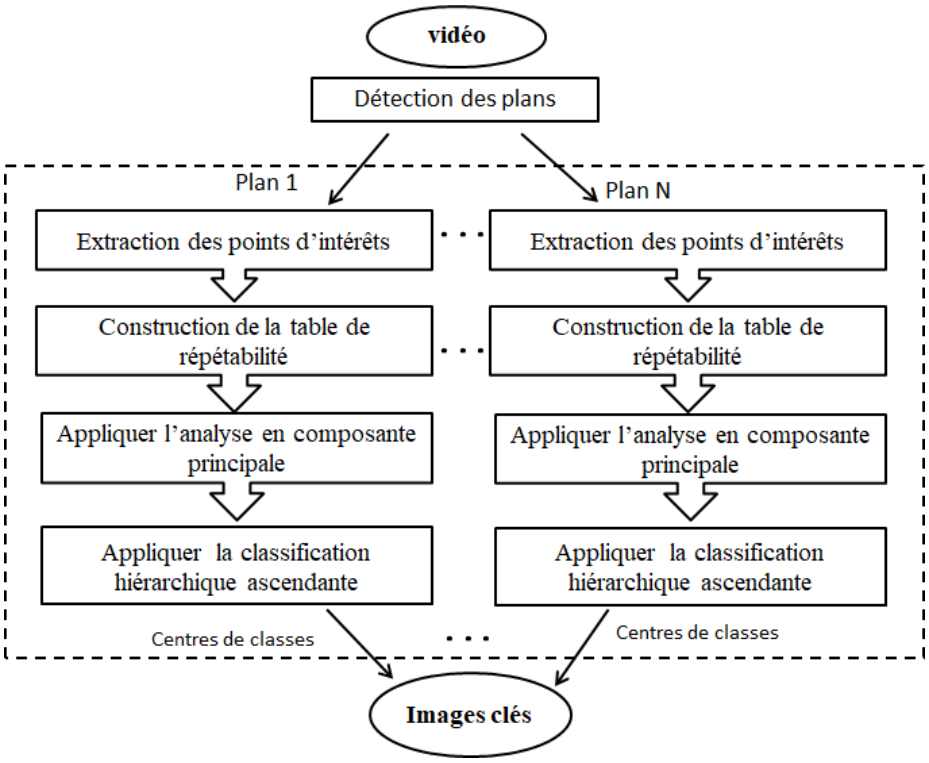

Figure 2.10. Processus général de la méthode EICCTR proposée

indépendamment des variations que peut subir une image [Gil, 2010]. L'équation 2.11 montre comment on calcule la répétabilité.

$$r_i = \frac{\mid R_i \mid}{min(n_1, n_2)} \tag{2.11}$$

Avec :

- R_i : représente le nombre de points qui sont répétés.

- n_1, n_2 : le nombre de points qui sont détectés dans les deux images.

Dans la littérature, plus précisément dans le cas des primitives locales qui sont les points d'intérêts, ce critère est considéré comme le plus fiable car il nous donne une indication sur les points d'intérêts répétés pour un ensemble d'images subissant plusieurs transformations. Dans notre travail, on l'applique pour chaque couple d'images appartenant à l'ensemble d'images constituants un plan de la vidéo. Dans

ce cas, une grande valeur de répétabilité indique une ressemblance au niveau du contenu des images, tans dis que une petite valeur indique un changement important au niveau de contenu. Dans ce contexte, on a construit une matrice pour chaque plan. La taille de la matrice carrée est N qui est le nombre d'images du plan.

Algorithme 1 : Algorithme de construction de la table de répétabilité

Données : RM : matrice de dimension N x N

N : nombre des images dans un plan

Résultat : RM : matrice de répétabilité remplie

initialisation;

pour $i \leftarrow 0$ **à** N **faire**

  pour $j \leftarrow i + 1$ **à** N **faire**

    // Appliquer l'algorithme d'appariement pour les deux images

    // Calculer la répétabilité entre les images i et j

    RM [i][j]= Répétabilité (i,j)

  fin

fin

Fin

Le calcul de répétabilité se fait selon la méthode de mise en correspondance MCIG présentée dans la section 2.2 et qui est basée sur la description locale par LBP et les invariants géométriques. Donc, si on considère un plan avec N images, cela va donner une matrice (diagonale supérieure) de répétabilité de taille N*N. L'algorithme 1 décrit le processus de construction de la table.

2.3.3 Classification des valeurs de répétabilité et sélection des images clés

De nombreuses techniques visant à partitionner une grande population en un ensemble de classes ou sous-groupes existent. Parmi ces techniques, on peut citer la classification ascendante hiérarchique (CAH). Le but de la CAH consiste à chercher les

	Image $_0$	....	Image $_i$	....	Image $_N$
Image 0	1		$RM_{i,j}$		
....		1			
Image $_j$			1		
....				1	
Images $_N$					1

Figure 2.11. Illustration de la table de répétabilité

individus qui peuvent être regroupés dans une même classe ou on parle de homogénéité intra-classe. Ces individus, qui sont assez dissemblables, sont très différents avec les éléments des autres classes ou on parle de hétérogénéité inter-classe. Le principe de cette méthode consiste à rassembler des différents individus selon un critère de ressemblance. Ce critère est défini au préalable selon le type de problème à traiter. Il est généralement exprimé en une matrice qui décrit les similarités entre chaque couple d'individus. La table de répétabilité obtenue vérifie les conditions d'application de cet algorithme. Mais, étant donné qu'elle est importante en termes de taille, cela peut engendrer un cout important et alourdir la complexité et le temps de calcul. C'est pour cette raison que l'application d'un algorithme de réduction de dimension s'impose comme une étape nécessaire. D'une façon générale, une étape de réduction de dimension est inévitable lorsque les données appartiennent à un espace de grande dimension. Ceci afin d'éviter les attributs redondants et sans aucune signification.
Vu que nous sommes dans le cas des séquences de vidéos où les images successives ne changent pas d'une façon significative (il y aura des données très proches qui sont la cause de certaines redondances). La perte d'information que peut causer la réduction de dimensionnalité de la table n'aura pas une influence significative sur la qualité des

résultats. Au contraire, elle minimise la quantité des variables similaires.
Dans ce contexte, nous avons appliqué l'analyse en composante principale (ACP). Sa principale idée est de réduire la dimension d'un jeu de données tout en gardant un maximum d'informations. Cette technique est capable de convertir un ensemble d'observations de variables éventuellement corrélées en un ensemble de valeurs de variables linéairement dé-corrélés. Le nombre final de variables est beaucoup moins important que le nombre initial ce qui nous permet d'avoir une représentation graphique sous forme de nuages de points. De plus, la classification dans une faible dimension est moins couteuse que celle dans une grande dimension. C'est ce qui a motivé l'utilisation de la réduction de dimension en utilisant l'ACP. Ainsi, l'algorithme ACP facilite la visualisation et l'interprétation des données et de réduire l'espace de stockage nécessaire. Cet algorithme permet de présenter le tableau de répétabilité sous forme de nuages de points affichés en 2 dimensions. Cette dimension a été choisie expérimentalement. En effet, après plusieurs tests nous avons remarqué qu'on peut retenir deux axes étant donnée qu'ils représentent presque 86 % de l'énetie totale.
Ces nuages de points obtenus seront divisés en classes en utilisant la Classification Ascendante Hiérarchique HAC. Mais, le problème qui persiste est quelle image choisir de chaque classe pour représenter l'image clé ? L'avantage de l'algorithme HAC est qu'il est simple, extrait automatiquement le nombre final de classes et nous donne le centre de chaque classe. Chaque centre de classe sera finalement représenté par une image clé. La figure 2.12 illustre le processus de la classification de la table de répétabilité.

2.3.4 Algorithme et complexité de la méthode d'extraction des images clés EICCTR

Dans cette section, nous déterminons la complexité de l'algorithme EICCTR. Pour ce faire, nous proposons de calculer la complexité de chaque étape de la méthode d'extraction des images clés EICCTR proposée. La construction de la table de

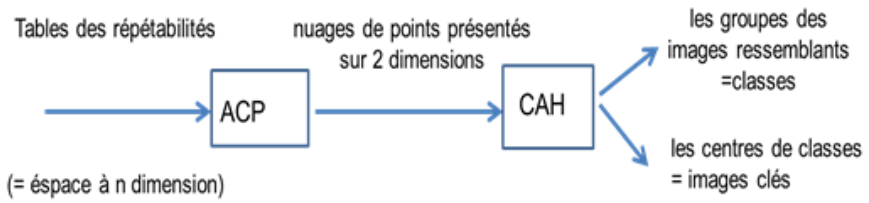

Figure 2.12. Illustration du processus de classification de la table de répétabilité.

répétabilité a une complexité $O(k^3)$. Elle permet de stocker les valeurs de répétabilité entre les couples des images dans une table de taille n*n. Le calcul de répétabilité entre chaque couple des images de la vidéo se fait à travers la méthode de mise en correspondance des points d'intérêts détectés. Sachant que k est le nombre de points d'intérêts détectés et n le nombre d'images par plan. L'application d'ACP et la CAH ont une complexité respectivement de l'ordre de $O(n^3/2)$ et $O(n^3)$. Ils permettent de réduire la dimension de la table puis classifier les ensembles de données afin d'extraire le centre de chaque classe.

2.4 Extraction d'images clés basée sur les graphes de répétabilité (EICGR)

Nous présentons dans ce paragraphe, une deuxième méthode d'extraction des images clés. Cette méthode est basée ,de même, sur la description locale par points d'intérêts et sur les mesures de répétabilité inter-images qui seront calculées aussi en utilisant la méthode d'appariement présentée dans la section 2.2. Cependant, elle réduit le nombre des images à traiter de la vidéo pour éviter le passage par l'étape de la réduction de dimension ACP et faciliter la classification. Dans le but de faciliter la sélection des images clés, la représentation des valeurs de répétabilité sera basée sur la notion de graphe. Le schéma général de la méthode ainsi que les différentes étapes seront détaillés dans les sections suivantes.

2.4.1 Description générale de la méthode d'extraction des images clés EICGR proposée

Les graphes, appelés aussi réseaux, sont considérés comme une modélisation naturelle qui peut être associée à un grand nombre de données lors de la résolution des problèmes réels où l'étude des entités n'est pas décrite uniquement par des attributs numériques ou encore qualitatifs mais en ajoutant des relations les reliant les unes aux autres. Ce type de données peut exister dans plusieurs domaines tels que par exemples : le domaine biomédical (voies métaboliques, régulation génique,..), les réseaux sociaux, l'informatique (réseaux de neurones, peer to peer, etc), l'ingénierie ou encore l'intelligence artificielle.

Les graphes à étudier dans ce type d'applications peuvent atteindre des centaines (ou même des milliers) de sommets à étudier ce qui nécessite des outils adaptés de fouille de données pour aider leur analyse et compréhension. Ces méthodes ont connu une grande progression ces dernières années tels que les méthodes de classification de nœuds de graphe, de visualisation de graphe,...

Dans ce contexte, nous avons proposé une méthode simple de sélection des images clés à partir d'un ensemble des images de la vidéo en s'inspirant du principe des algorithmes à plus court chemin. Cette méthode a donné de bons résultats en comparaison avec des méthodes existantes dans la littérature. Cependant, pour certaines vidéos ou le contenu est relativement stable et les images sont très similaires, elle peut générer des images clés contenant des redondances. Cet inconvénient nous a encouragé à proposer une amélioration et ce en introduisant la classification de graphe par maximisation du terme de modularité lors de l'étape de sélection des images clés. Ces deux variantes seront bien détaillées dans les sections suivantes.

La méthode de sélection des images clés basée sur la description locale et utilisant les deux algorithmes ACP et CAH décrite précédemment a donné de bons résultats face à des méthodes existantes dans l'état de l'art malgré le passage par

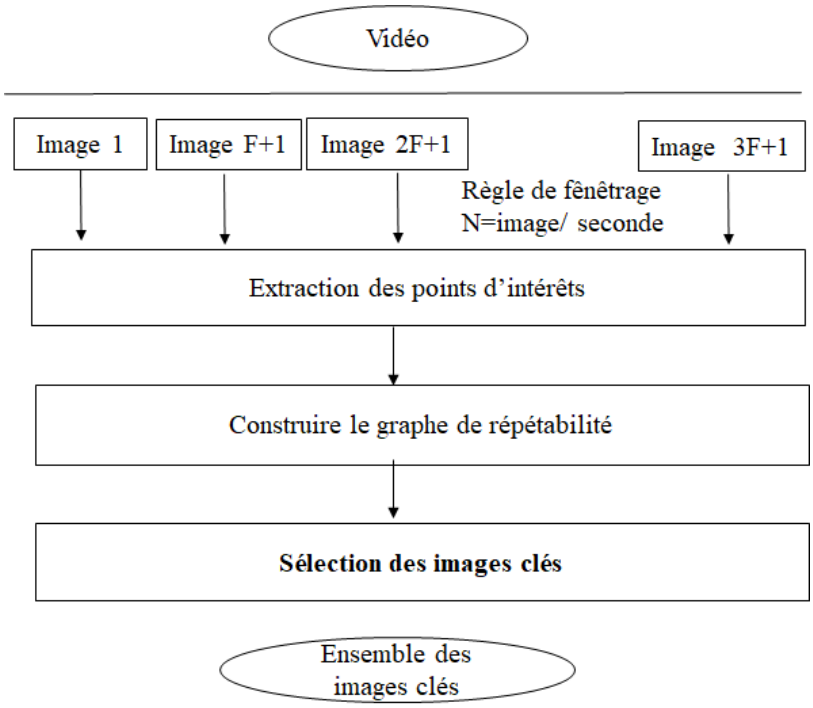

Figure 2.13. Processus général de la méthode proposée pour l'extraction des images clés en se basant sur la représentation graphique

une réduction de dimension qui peut causer une perte de données et qui possède un cout relativement élevé en terme de complexité. Ceci prouve l'efficacité de la description locale par points d'intérêt dans la génération du résumé de vidéo. Dans le souci d'avoir des résultats meilleurs (des images clés plus fidèle aux vidéos d'origine et minimiser la complexité de calcul de certaines étapes), nous avons proposé une deuxième méthode de sélection d'images clés basée sur les graphes. Sachant que le nombre des points d'intérêts existants et le nombre des images qui composent la vidéo est important, dans cette méthode, l'analyse des descripteurs locaux s'effectue seulement sur un nombre de candidats pas toutes les images de la vidéo. Ces candidats ne sont pas choisis aléatoirement mais en se basant sur une règle de fenêtrage. Puis, la table de répétabilité , construite uniquement pour les images candidates, pourra être

représentée par un graphe. L'algorithme de sélection des images clés sera effectué selon deux alternatives. Ces deux alternatives seront expliquées en détails dans les paragraphes suivants.

2.4.2 Construction du graphe de répétabilité

Généralement, la vidéo contient un nombre important d'images. Ces images sont affichées à une fréquence de 25 à 30 images par seconde. Pour les vidéos génériques, les scènes changent normalement lentement. Ceci permet un échantillonnage de la vidéo d'entrée sans avoir un impact significatif sur le résultat du résumé.

- Génération des images candidates

Dans le but d'éviter la comparaison des images qui sont presque similaires en terme de contenu et afin de minimiser le cout de traitement de ces images, nous avons choisi de sélectionner un certain nombre d'images parmis l'ensemble des images d'une vidéo. L'ensemble de ces images seront appelées ensemble des images candidates (CS). La technique utilisée dans la sélection de ces images est celle du fenêtrage. La première image de chaque plan de vidéo est insérée par défaut dans le (CS). Ceci est dans le but de garantir que chaque plan sera au moins représenté par une image clé. Ensuite, en suivant la règle de fenêtrage, le reste des images candidates sera inclut dans le CS. La fenêtre que nous avons définit est de taille F. Puis les images aux positions F+1, 2F+1, 3F+1 seront extraites pour être analysées ultérieurement. La valeur de F a été fixée expérimentalement pour la valeur de la FPS (frame par seconde) vu que dans une seule seconde on ne peut pas trouver une variation significative dans le contenu des images consécutives [Dang, 2015].
L'algorithme 2 décrit le processus de sélection des images candidates.

Algorithme 2 : Algorithme de sélection des images candidates

Données : Video V=f1,f2,. . . ,fn

Résultat : cs

initialisation;

fps := V.getFPS()

i := 1

tant que $i < n$ **faire**

    cs.add(fi);

    i = i + fps;

fin

Fin

- Construction de graphe

Dans cette étape, l'extraction des descripteurs est effectuée seulement pour les images candidates de chaque plan et non pas pour la totalité des images du plan comme la méthode MCIG présentée dans le paragraphe 2.3.1. D'ou, au lieu de construire une table de répétabilité pour toutes les images de chaque plan, on l'a construit uniquement pour les images candidates. Comme nous avons cité précédemment, la répétabilité [Parks, 2010] est le critère le plus efficace pour le jugement de la ressemblance entre les images subissant différentes transformations en utilisant la description locale par points d'intérêts (ces images sont dans notre cas les images candidates de la vidéo). L'algorithme 3 montre comment on construit la table de répétabilité pour l'ensemble des images candidates de chaque plan. La table résultante est sous la forme d'une matrice d'adjacence. C'est une matrice de dimension N*N avec N est le nombre d'images candidates sélectionnées, dont les éléments non diagonales, notées r_{ij}, représentent la répétabilité entre les sommets i et j. L'étape suivante consiste à sélectionner les images clés à partir de cette table. Le but de simuler notre problème au problème de graphe est d'assurer la qualité du résultat généré. En effet, cette table de répétabilité

Algorithme 3 : Algorithme de construction de la table de répétabilité pour chaque plan

Données : T : matrice de dimension N x N

N : nombre de (CS) dans un plan

Résultat : RM : matrice de répétabilité remplie

initialisation;

pour $i \leftarrow 0$ **à** N **faire**

  pour $j \leftarrow i + 1$ **à** N **faire**

    // Appliquer l'algorithme d'appariement pour les deux images candidates

    // Calculer la répétabilité entre les images i et j

    T [i][j]= Répétabilité (i,j)

  fin

fin

Fin

pourra être représentée par un réseau R(X,E,d). X désigne l'ensemble des sommets de R, E est l'ensemble des arcs reliant les sommets de X et d est l'application distance définie comme suit :

$$d : E \rightarrow R \tag{2.12}$$

$$e \rightarrow d(e)\, qui\, est\, la\, distance\, de\, l'arc\, e$$

Relativement à notre contexte, X désigne l'ensemble des images candidates avec $|X| = N$. Ainsi, chaque image i est un sommet i du graphe G(X, E). E est l'ensemble des arcs reliant ces images et d est la répétabilité entre chaque paire d'images. Chaque sommet i est relié aux sommets i+1,. . . , N. Le graphe G(X, E) obtenu est complet, sans circuit, possède une source (l'image 1) qui est également une racine et un puits (l'image N). Notons que les images candidates sont numérotées relativement dans leur ordre chronologique. On considère l'exemple suivant pour 4 images candidates 1,2, 3 et 4, r_{ij} est la répétabilité associée aux images i et j. Le sens de flèche est relatif à l'ordre chronologique. On construit ainsi, pour les images candidates de chaque plan

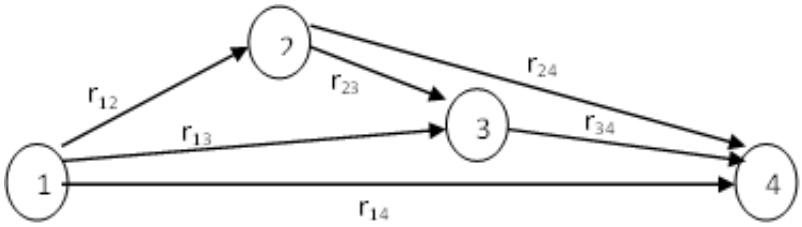

Figure 2.14. Exemple Illustratif d'une représentation graphique de la table de répétabilité

un réseau à partir duquel seront sélectionnées les images qui formeront le résumé de la vidéo.

2.4.3 Sélection des images clés

Pour la sélection des images clés qui forment les résumés statiques, nous avons proposé deux alternatives :

- La première alternative est basée sur la mesure de répétabilité minimale. Elle est inspirée à partir du principe des algorithmes du plus court chemin. (EICGR-1)
- La deuxième alternative est basée sur la classification du graphe par calcul de modularité. (EICGR-2)

2.4.3.1 Sélection de la répétabilité minimale (EICGR-1)

Le principe de la méthode proposée est inspiré du principe des algorithmes du plus court chemin. Cela consiste à résoudre le problème en cherchant parmi tous les chemins possibles, vers l'objectif, celui qui donne le plus petit coût. Dans notre cas, le coût est relatif à la valeur de répétabilité. Le graphe est orienté et exige que les sommets consécutifs soient connectés par une arrête orientée appropriée puisque les valeurs de répétabilité sont ordonnées dans un sens chronologique. Donc, on doit commencer par chercher pour chaque graphe la valeur minimale de répétabilité. En effet, la valeur de répétabilité traduit la ressemblance entre les images en termes de contenu. Donc, une valeur minimale de celle-ci traduit la plus faible ressemblance entre les images et inversement. L'idée initiale consiste à chercher le sommet, dont l'arrête sortante possède le coût minimum, cette arrête lui conduit vers le sommet le moins ressemblant en terme de contenu (répétabilité minimale de la table). Cette valeur minimal doit être à un seuil prédéfinit (=0,2). Ce seuil a été fixé après plusieurs expérimentations. Son choix a été relativement strict pour garantir l'extraction des images clés ayant un contenu très hétérogène. Une fois le sommet traité, on passe au sommet suivant. Donc, pas de retour en arrière, ce qui est bénéfique pour l'élimination de la redondance. Ainsi, à partir de ce raisonnement, nous avons procédé à sélectionner les images clés à partir de la table de répétabilité

Algorithme 4 : Sélection d'images clés pour la méthode EICGR-1

```
Données :
T[N][N]; // Matrice de répétabilité de dimension N x N avec N nombre de CS
KS= Ø ; // Ensemble d'images clés
min; // Répétabilité minimale de la matrice de répétabilité
i=j=0;
S=0,2;
Résultat : KS;
initialisation;
tant que j < N faire
    tant que i < N faire
        si T[i][j]==min alors
            ajouter i dans KS;
            si (min < S)
             alors
              | ajouter j dans KS;
            fin
            i=j;
        sinon
          | j++;
        fin
    fin
    i++;
    j=i;
fin
Fin
```

2.4.3.2 Classification des valeurs de répétabilité par maximisation de la modularité (EICGR-2)

Nous avons proposé cette alternative EICGR-2 dans le but d'améliorer la première EICGR-1 présentée dans le paragraphe précédent. La figure 2.15 montre le processus général de la méthode EICGR en utilisant cette alternative.

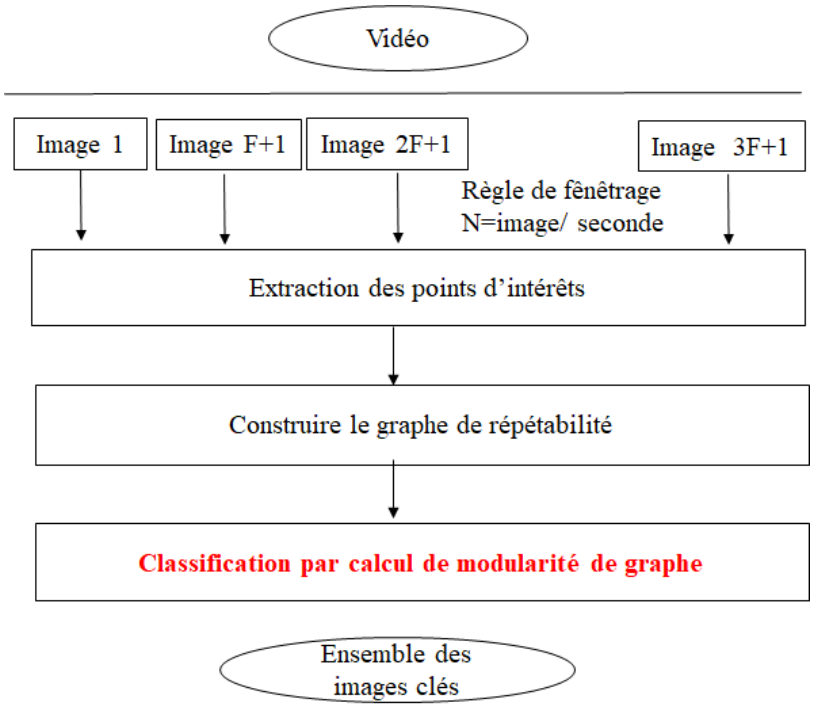

Figure 2.15. Processus de sélection des images clés basé sur la classification par calcul de modularité

Dans cette alternative, pour la sélection des images clés nous allons utiliser la classification automatique qui est considérée comme méthode de classification non supervisée. Elle permet le partitionnement d'un ensemble d'observations sous forme de classes. En effet, la classification automatique conduit à la partition d'une population initiale en un ensemble de groupes disjoints comme illustré dans la figure 2.16,

de sorte que deux individus appartenant à un même groupe auront entre eux un maximum d'affinité et inversement deux individus appartenant à des groupes différents auront un minimum d'affinité. Ceci est effectué selon un critère bien défini selon le contexte. Dans le contexte de description locale par points d'intérêts, la répétabilité est retenue comme critère de similarité en termes de contenu.

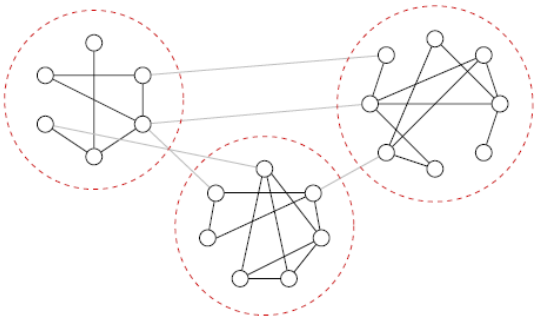

Figure 2.16. Illustration du principe de partitionnement de graphe en communautés.

En fait, le principe de classification de graphes consiste à extraire des groupes de sommets, appelés communautés. Ces sommets sont connectés de façon dense et partagent essentiellement un minimum de caractéristiques avec les sommets appartenant aux restes de communautés [Fortunato, 2010]. Le but principal de la décomposition est de faciliter l'exploration et la compréhension du réseau. Le fait de se focaliser sur un nombre réduit de classes (groupes d'images) permet à l'utilisateur de mieux extraire les caractéristiques de chacune.

Dans la littérature, plusieurs méthodes ont été développées pour la classification de sommets d'un graphe [Rossi, 2012][Fortunato, 2010]. Elles sont généralement basées sur la définition d'une mesure de similarité reliant les sommets (peut s'appuyer sur un plongement du graphe dans son espace euclidien), ou encore à travers des méthodes génératives supposant que le graphe peut être généré en se basant sur un modèle

aléatoire ou les densités inter- communautés et intra-communautés sont différents comme dans [Daudin,2008] [Zanghi,2008], ou encore elle peuvent être basées sur une optimisation d'un critère de qualité de la classification, tel que l'exemple de la populaire modularité qui a été introduit dans [Newman, 2004]. Plusieurs revues telles que dans ([Fortunato, 2010], [Schaeffer, 2007]) ont donné un panorama complet des différentes méthodes de classification de sommets composants un graphe. La mesure de modularité a été introduite pour la classification de graphes et a montré des résultats performants. Citons l'exemple de ([Agarwal, 2008] et [Rossi, 2012]).
En effet, la modularité permet de guider la recherche de la partition P. Plus spécifiquement, la modularité permet de mesurer pour chaque partition P possible une valeur M(P)de modularité. celle ci fournit un indice sur la qualité de la partition générée. La maximisation de cette fonction M permet l'identification de la meilleure structure de communautés dans le réseau donné.
Deux catégories d'approches sont largement étudiées :
– Les approches agglomératives : appelées aussi ascendantes, selon lesquelles on part de la partition atomique (ensemble des singletons), et on fusionne deux communautés à chaque itération. Les communautés à fusionner sont celles qui promettent une modularité maximale. Un exemple de cette catégorie est donné dans [Newman, 2004].

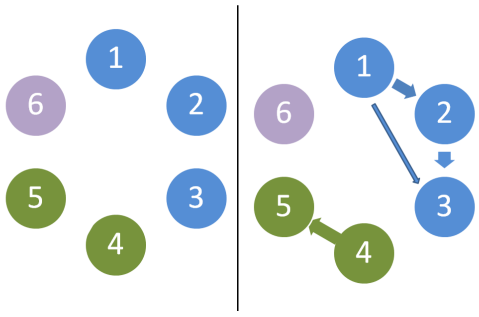

Figure 2.17. Exemple Illustratif du principe des approches agglomératives

– Les approches divisives : appelées aussi descendantes, selon lesquelles on part d'un graphe entier. A chaque itération, on cherche à scinder une communauté parmi celles

existantes en deux de sorte à maximiser la fonction de modularité. Un exemple de cette catégorie est donné dans [Rossi, 2012].

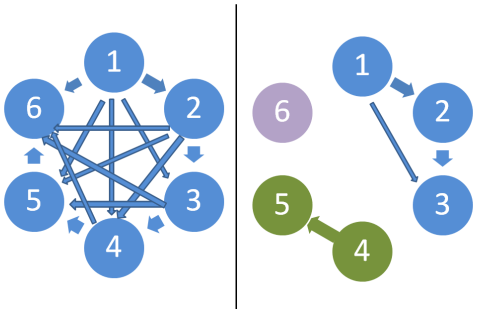

Figure 2.18. Exemple Illustratif du principe des approches divisives

Notre travail s'inscrit dans le contexte des approches divisives. Ainsi, nous avons adapté le graphe de répétabilité à ce contexte et ce pour trouver la partition optimale maximisant le critère de modularité en relation avec la répétabilité qui est le critère essentiel reliant les sommets du graphe (qui représentent les frames candidates). Pour ceci, nous avons étendu le problème de maximisation de modularité pour faciliter la sélection des images clés.

- Génération des images candidates

La technique utilisée dans la sélection de ces images est celle décrite précédemment (paragraphe 2.4.2) : technique de fenêtrage. Cependant dans cette alternative, seulement la première image de chaque de vidéo est insérée par défaut dans le (CS). Ensuite, en suivant la règle de fenêtrage pour toute la vidéo, le reste des images candidates sera inclut dans le CS. De ce fait, le découpage de la vidéo en plans n'est plus utile.

- Construction du graphe de répétabilité

Dans cette partie le graphe est construit en utilisant toutes les images appartenant à l'ensemble des images candidates de la vidéo entière et les valeurs de répétabilité entre

chaque couple d'images. De même que dans le paragraphe 2.4.2, les images candidates sont représentées par les sommets du graphe. Chaque sommet est connecté à tous les sommets qui le suivent dans l'ordre chronologique. Par la suite, les arêtes connectant chaque deux sommets seront pondérées par la valeur (1- valeur de répétabilité) étant donné que le poids correspond à la distance reliant les sommets. Dans le cas général, une faible valeur de poids signifie que les sommets se ressemblent plus, inversement en cas de faible valeur de répétabilité signifie que les sommets (images) se ressemblent moins. C'est pour cette raison qu'on a introduit la valeur (1-répétabilité). On aura comme résultat un graphe complet orienté qui connecte tous les sommets (images candidates de toute la vidéo)dans un ordre chronologique.

- Classification par maximisation de modularité

Dans le graphe résultant de la classification, les arcs doivent être groupés en intra-classe (les sommets qui appartiennent au même groupe) et inter- classe (les sommets qui appartiennent au groupes différents). Le principe de la classification d'un graphe de similarité G en se basant sur la maximisation de modularité est de préserver les arcs intra-classe et supprimer les arcs inter-classe. Le principe est d'enlever certains arcs en fonction de la différence entre les poids d'arcs, jusqu'à ce qu'il n'y ait pas une amélioration dans la valeur de modularité du graphe. En effet, la valeur la plus grande de modularité indique une meilleure classification [Agarwal, 2008] [Schaeffer, 2007]. Dans ce contexte, on a modifié la fonction de poids pour qu'elle s'adapte mieux au principe de répétabilité (1-répétabilité).

La modularité M (c1, c2, ..., ck) pour la classification de graphe pour un nombre de k classes c1, c2, ..., ck est défini comme suit:

$$M(c_1, c_2, ..., c_k) = \sum_{i=1}^{k} \delta_{i,i} - \sum_{i \neq j} \delta_{i,j}, \tag{2.13}$$

$$Etant\, \delta_{i,j} = \sum_{\{u,v\} \in E, v \in c_i, u \in c_j} w(v,u)$$

Notons que chaque arête $v, u \in E$ n'est incluse qu'au plus une fois dans le calcul. Plus la valeur de répétabilité est élevée, plus la classification est meilleure.

Algorithme 5 : Sélection d'images clés pour la méthode EICGR-2

Données :

G, E, W, T [N][N]

Résultat : Clusters c1, c2, ..., ck

initialization;

répéter

Sélectionner les arcs de plus grande valeur;

Eliminer ces arcs du G;

Trouvez les composants connectés du G;

Calculer la modularité (M);

jusqu'à

Aucune amélioration de la modularité sur deux itérations successives;

Obtenir les clusters individuels à partir du G final représentés par des sous graphes disjoints;

Fin

Les autres composants connectés du graphe final après la fin de l'élimination des arêtes représentent les classes individuelles. Sachant que T est la table de répétabilité avec N nombre de (CS). L'algorithme 5 résume les différentes étapes du processus de classification du graphe en utilisant la maximisation de la valeur de modularité.
L'image candidate qui est plus proche du centre de chaque classe est considérée comme image clé. Enfin, les images clés sont organisées dans l'ordre chronologique pour rendre le résumé produit plus compréhensible.

2.4.4 Algorithme et complexité de la méthode EICGR

Pour la méthode d'extraction des images clés basée sur les graphes de répétabilité EICGR, nous avons proposé deux alternatives pour la sélection des images clés. On propose dans cette section de déterminer la complexité de chacune de ces alternatives.

Pour EICGR-1, la complexité de la fonction de génération des images candidates est de l'ordre O(C) avec $C = N/K$ tel que N = nombre d'image candidats par plan et $K = FPS$. La fonction de construction de table de répétabilité est évaluée à $O(M^3)$ sachant que M est le nombre de points d'intérêts détectés, les fonctions de recherche de la répétabilité minimale et de sélection des images clés sont chacune de l'ordre $O(C^2)$. Pour EICGR-2 la complexité de fonction de génération des images candidates est de l'ordre $O(C)$ avec $C = N/K$ tel que N = nombre d'images candidates de la vidéo et $K = FPS$, la fonction de construction de graphe de répétabilité est $O(C^2)$ et celle de la classification par maximisation de modularité est de l'ordre $O(C^2 \, log \, C)$.

Conclusion

Au cours de ce chapitre, nous avons introduit les méthodes proposées pour la génération du résumé de vidéos statique. Pour ce faire, nous nous sommes basés sur la description locale par points d'intérêts. Cette primitive est une bonne alternative pour une caractérisation robuste des images de la vidéo vu sa capacité en termes d'invariance face aux divers changements, bien que parmi les méthodes d'extraction des images clés existantes dans la littérature, il existe très peu qui ont tiré profit de cette description locale. Puisque le résultat dépend non seulement de la robustesse de la méthode mais aussi des primitives extraites, nous avons effectué une étude comparative des différents détecteurs pour choisir le plus adapté à notre contexte. Le critère de répétabilité est le plus populaire pour évaluer les détecteurs des points d'intérêts. Le calcul de répétabilité nécessite une méthode de mise en correspondance pour calculer le nombre de points appariés après changements. Pour ce faire, nous avons proposé une méthode qui comprend les deux dernières étapes du processus d'extraction et qui sont : description et mise en correspondance.

Chapitre 3

EXPÉRIMENTATIONS ET ÉVALUATION

Introduction

Dans ce chapitre, nous allons évaluer les différentes méthodes proposées dans ce travail de thèse : la méthode d'extraction des caractéristiques locales MCIG puis celles d'extraction des images clés de vidéos EICCTR, EICGR-1 et EICGR-2. Pour chacune de ces méthodes, nous évaluons tout d'abord la qualité des résultats en utilisant des métriques d'évaluation subjectives. On se base pour cela sur un ensemble de données fournies par les bases : vérité terrain. Ensuite, nous allons effectuer une évaluation objective de ces méthodes. Cette évaluation comportera une étude comparative avec les méthodes existantes les plus citées dans la littérature. Enfin, pour s'assurer davantage de l'efficacité des méthodes proposées, nous avons projeté les résultats obtenus dans un système de recherche de vidéo par le contenu. Toutes les expérimentations ont été implémentées sous Microsoft Visual C++ 2010, en utilisant la bibliothèque OpenCV 2.4.3, sur un PC de processeur Intel Core (TM) i5, CPU 2.50GHZ et de 6GB de RAM.

3.1 Mise en correspondance des points d'intérêts

Tout au long de cette section, nous allons expliquer le protocole d'évaluation que nous avons suivi afin de prouver l'efficacité de la méthode de mise en correspondance MCIG proposée.

3.1.1 Protocole d'évaluation

La fiabilité d'un tel protocole repose sur deux éléments principaux. Tout d'abord, l'adoption de métriques d'évaluation adéquates qui estiment la qualité des résultats. Par la suite, le choix de bases d'images appropriées prises dans des conditions variées afin de garantir une invariance face aux différents changements possibles que peut subir une image.

3.1.1.1 Base des images

Lors de nos expérimentations, nous avons utilisé deux bases d'images, ces deux bases sont généralistes et elles sont communément utilisées dans la littérature pour l'évaluation des méthodes de mise en correspondance.

- "Zubud" : elle est librement disponible sur internet [Shao, 2003]. Elle contient plus de 1005 images concernant bâtiment-ville de Zurich. Les images de cette base sont de taille 640*480 pixels ou 320*240 pixels. Elles sont prises à partir des angles de vue aléatoires, sous occlusion, des conditions variables d'échelle et de luminosité.

- "Oxford" : librement accessible sur internet. Elle contient des séquences d'images subissant plusieurs types de transformations. Ensemble de séquences avec chacune 6 images montrant différentes scènes structurées et texturées. Chaque séquence montre différentes transformations d'images : Ces transformations comprennent le changement d'angle de vue, le zoom, le flou et la rotation. Les images de cette base sont de tailles 800 x 640 pixels.
  [http://www.robots.ox.ac.uk/ vgg/research/affine/].

3.1.1.2 Métriques d'évaluation

Les différents tests effectués permettent la validation et la mise en avant des avantages et des faiblesses de la méthode de mise en correspondance MCIG proposée. Il est très important aussi d'avoir à la fois des critères qualitatifs et quantitatifs qui permettent de juger les performances de la méthode de mise en correspondance proposée. Pour ce faire, nous avons utilisé l'ensemble des métriques suivantes :

- **Le nombre de points appariés**

C'est le nombre d'appariements qui résultent de la méthode de mise en correspondance proposée. Ce critère caractérise l'aspect quantitatif des résultats de la mise en correspondance.

- **La précision**

$$Présion = \frac{Nombre\,de\,bons\,appariements}{Nombre\,d'appariements\,trouvés}$$

Ce taux permet d'évaluer la qualité de la mise en correspondance et de fortifier la pertinence du descripteur local.

- **Temps d'exécution**

En plus de la précision et le taux d'appariements, il est important de mesurer le temps de calcul pour garantir la qualité de la méthode proposée par rapport aux autres méthodes de la littérature.

3.1.2 Résultats expérimentaux

Nous allons comparer notre méthode aux trois méthodes de mise en correspondance SIFT [Lowe, 2004], SURF [Bay, 2008] et PW-MATCH [Kingsbury, 2010] . Cette comparaison aura pour objectif de montrer l'apport du descripteur local LBP pour la description des points d'intérêts, ainsi que l'apport des contraintes spatiales basées sur les invariants géométriques lors de la phase de mise en correspondance. En effet, nous avons choisi SIFT et SURF car il a été démontré dans la littérature [Brochier,

2010] [Awad, 2016], que ces deux méthodes restent des références dans le processus de description locale par points d'intérêts. Pour PW-MATCH aussi il a montré une bonne performance [Kingsbury, 2010], en plus du fait qu'il a été basé sur les contraintes spatiales. Pour ce faire, nous étudierons le taux d'appariement ainsi que la stabilité selon différentes transformations que peut subir une image.

3.1.2.1 Exemples de résultats

Pour une évaluation réussite, plusieurs types de transformations doivent être étudiées : le changement de luminosité, le changement de point de vue petit angle et grand angle, le couplage rotation/changement d'échelle.

Dans ce contexte, nous allons commencer par présenter des exemples de résultats de mise en correspondance pour quelques images appartenant aux deux bases (image initiale et celle qui subit la transformation) puis nous allons travailler sur des séquences d'images composées d'une image originale subissant un ensemble de transformations dégradées du plus petite vers la plus grande afin d'étudier la robustesse de la méthode MCIG face à différentes transformations.

Figure 3.1. Exemple de résultat d'appariement lors d'une rotation

Les figures 3.1, 3.2 et 3.3 montrent les résultats obtenus après une mise en correspondance effectuée par la méthode proposée MCIG entre des images références (celles à gauche) et d'autres subissant des transformations (celles à droite). Ces figures sont suivies respectivement par les tableaux 3.1, 3.2 et 3.3. Ces tableaux comparent ces résultats avec celles de SIFT, SURF et PW-MATCH en termes de nombre des points appariés et de précision.

	SIFT	SURF	PW-MATCH	MCIG
Points appariés (a,b)	108	66	90	92
Précision	0.81	0.84	0.84	0.86

Table 3.1. Tableau comparatif des résultats obtenus pour les images de la figure 3.1 lors de l'appariement des deux images avec la méthode proposée MCIG ainsi que SIFT, SURF et PW-MATCH

(c) (d)

Figure 3.2. Exemple de résultat d'appariement obtenu lors d'un changement d'angle de vue

	SIFT	SURF	PW-MATCH	MCIG
Points appariés (c,d)	171	96	147	111
Précision	0.83	0.79	0.85	0.88

Table 3.2. Tableau comparatif des résultats obtenus pour les images (c) et (d) lors de l'appariement des deux images avec la méthode proposée ainsi que SIFT, SURF et PW-MATCH

Figure 3.3. Exemple de résultat d'appariement d'un couple d'images de la même scène

3.1.2.2 Evaluation selon différentes transformations

Nous présentons dans cette section une évaluation qualitative des résultats de la méthode proposée sur les différentes séquences de test. Une image de référence est utilisée (image 1 à gauche de chaque séquence), celle-ci sera appariée avec les images qui la suivent respectivement dans chaque séquence. Chacune des séquences d'images est dédiée à étudier un type de transformation.

	SIFT	SURF	PW-MATCH	MCIG
Points appariés (e,f)	580	270	469	320
Précision	0.81	0.91	0.84	0.90

Table 3.3. Tableau comparatif des résultats obtenus pour les images (e) et (f) lors de l'appariement des deux images avec la méthode proposée ainsi que SIFT, SURF et PW-MATCH

a - Changement d'angle de vue :

Figure 3.4. Image référence "Graffiti" qui subit des transformations de changement d'angle de vue (image 1, ..., image 6)

Dans ce paragraphe, nous présentons les résultats de la méthode d'appariement proposée pour une séquence d'images avec un changement d'angle de vue dégradé de l'angle petit vers le plus grand. Pour ce faire, nous avons pris la séquence de test "Graffiti" de la figure 3.4. Dans cette séquence, l'image initiale (image 1) subit un changement d'angle de vue. Ce changement s'accentue progressivement de l'image 2 vers l'image 6.

Nous présenterons respectivement dans les figures 3.5 et 3.6 les résultats de nombre d'appariements et de précision pour des changements d'angle de vue. Ces changements s'accentuent respectivement du changement numéro 1 (entre l'image 1 et l'image 2) vers le changement numéro 5 (entre l'image 1 et l'image 6).

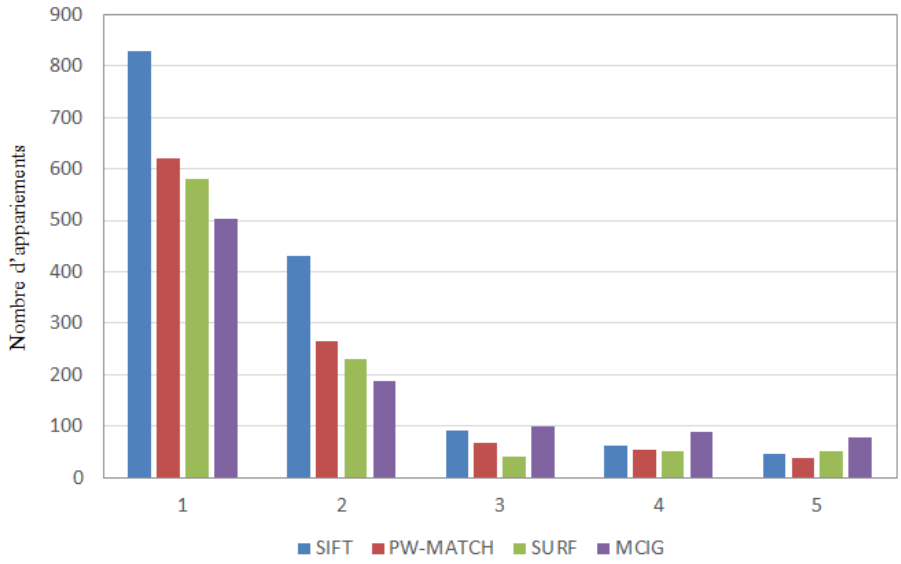

Figure 3.5. Résultats comparatifs en termes de nombre d'appariement trouvés lors des transformations de changements d'angle de vue successives pour l'image "Graffiti"

Pour ce type de transformation, la méthode MCIG proposée surpasse nettement les méthodes SIFT et PW-MATCH en terme de précision. En comparaison avec la méthode SURF, elle présente des résultats presque similaires pour les petits changements d'angles de vue et meilleurs pour les changements les plus importants. De plus, on peut noter que la courbe de précision de la méthode MCIG est la plus constante lors des changements: elle décroit moins rapidement d'un changement à un autre plus important, contrairement aux courbes des autres méthodes qui sont nettement décroissantes.

En ce qui concerne le nombre d'appariements, les résultats de la méthode MCIG proposée décroissent aussi de manière plus constante que les autres méthodes, d'un changement à un autre plus important, bien que ce nombre n'est pas le plus important pour la totalité des transformations.

Ces deux figures prouvent la stabilité de la méthode MCIG face aux changements

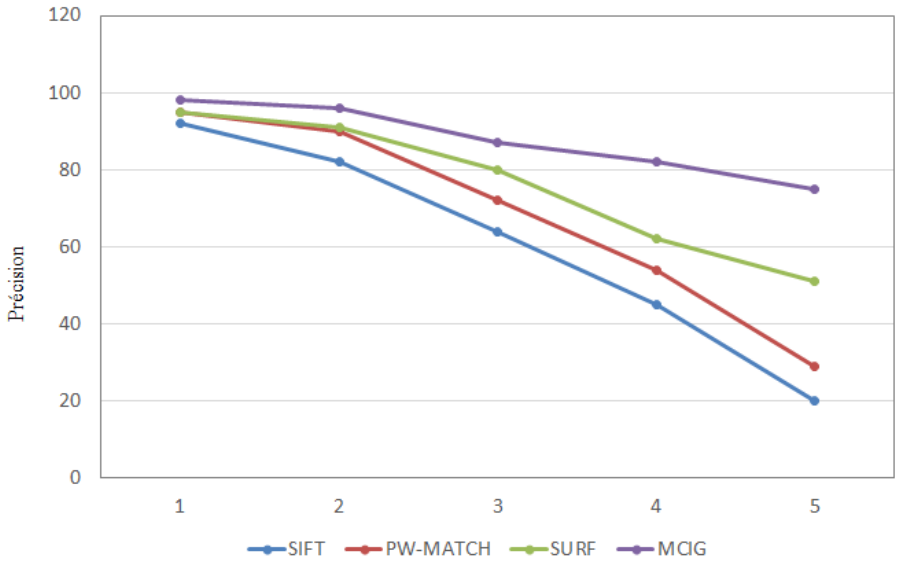

Figure 3.6. Résultat en termes de précision lors d'un changement progressif d'angle de vue

d'angles de vue et confirment que l'ajout des contraintes spatiales en se basant sur les invariants géométriques minimise considérablement le taux des faux appariements.

b - Couplage rotation et changement d'échelle

Dans ce paragraphe, nous montrons la robustesse de la méthode MCIG proposée sur un ensemble d'images avec un changement en couplage de rotation et changement d'échelle. Pour ce faire, nous avons pris la séquence de test "boat" présentée dans la figure 3.7. Dans cette séquence, l'image initiale subit un changement progressif de rotation et changement d'échelle (qui s'accentue de l'image 2 vers l'image 6).

Nous présenterons dans les figures 3.8 et 3.9 les résultats, en termes de nombre d'appariements et de précision, pour des changements en couplage (Rotation +changement d'échelle) qui s'accentuent du changement numéro 1 (entre l'image 1 et l'image 2) vers le changement numéro 5 (entre l'image 1 et l'image 6).

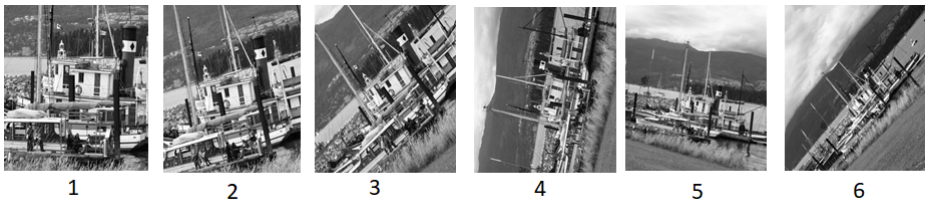

Figure 3.7. Image référence " boat" qui subit des transformations en couplage (Rotation+changement d'échelle)

D'après la figure 3.9, la méthode MCIG proposée présente un meilleur résultat en

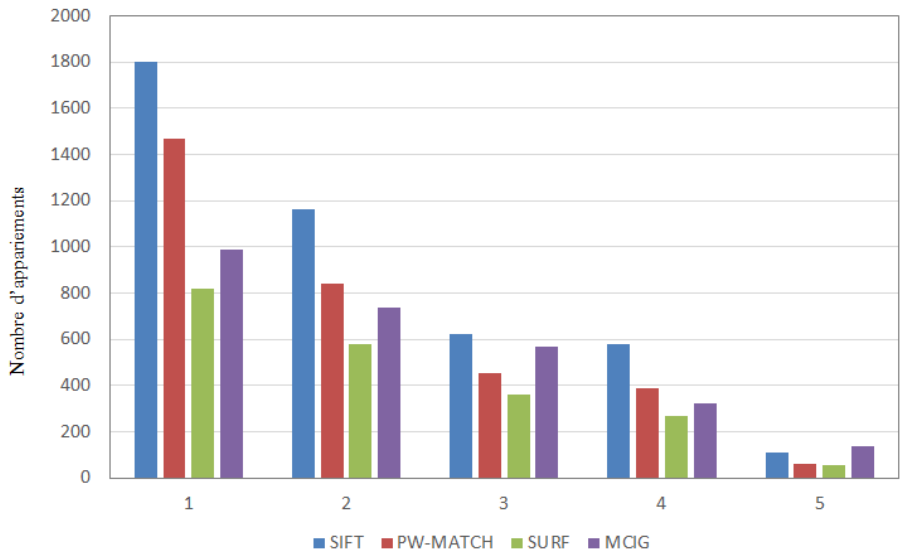

Figure 3.8. Tableau comparatif en termes de nombre d'appariements trouvés lors des transformations en couplage (rotation + changement d'échelle)

termes de précision que les autres méthodes SIFT, SURF et PW-MATCH. Ce résultat est plus clair pour les transformations les plus importantes (dans le cas où la valeur de rotation et de changement d'échelle augmente). Ce qui montre davantage la stabilité des résultats trouvés indépendamment de la transformation (petite ou grande).

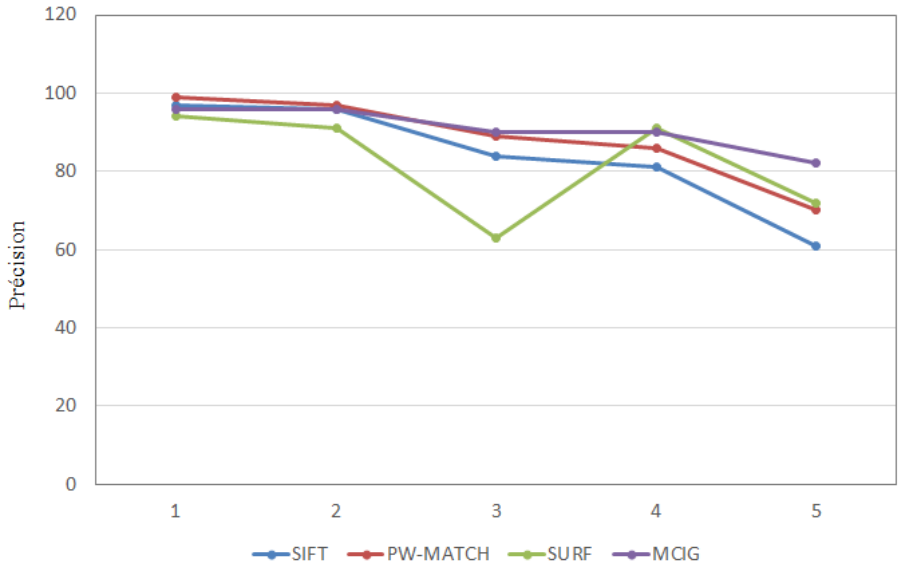

Figure 3.9. Résultat en termes de Précision lors des transformations en couplage (rotation + changement d'échelle)

En ce qui concerne le résultat en termes du nombre d'appariements, il est meilleur que le SURF pour toutes les transformations. Mais, il est plus faible que le SIFT et PW-MATCH surtout pour les petites (premières) transformations.
Ces résultats confirment la réussite des deux étapes celle d'extraction du descripteur LBP et celle de mise en correspondance par invariants géométriques à minimiser le nombre de faux appariements, tout en montrant une meilleur stabilité face aux changements en couplage (rotation et changement d'échelle).

c - Changement de luminance

Cette section présente la robustesse de la méthode d'appariement proposée à un changement de luminosité. Pour ce faire, nous avons pris la séquence de test "cars", présentée dans la figure 3.10, dans laquelle l'image initiale subit un changement progressif de luminosité (qui s'accentue de l'image 2 vers l'image 6).

Figure 3.10. Image référence "cars" qui subit un ensemble de changement de luminance progressive

Le graphe de la figure 3.11 montre le résultat obtenu en termes de nombre d'appariements lors de la comparaison de la méthode proposée avec les algorithmes SIFT et SURF pour des changements de luminosité qui s'accentuent du changement numéro 1 (entre l'image 1 et l'image 2) vers le changement numéro 5 (entre l'image 1 et l'image 6). Ce graphe est suivi de la figure 3.12 qui présente une courbe comparative en termes de précision.

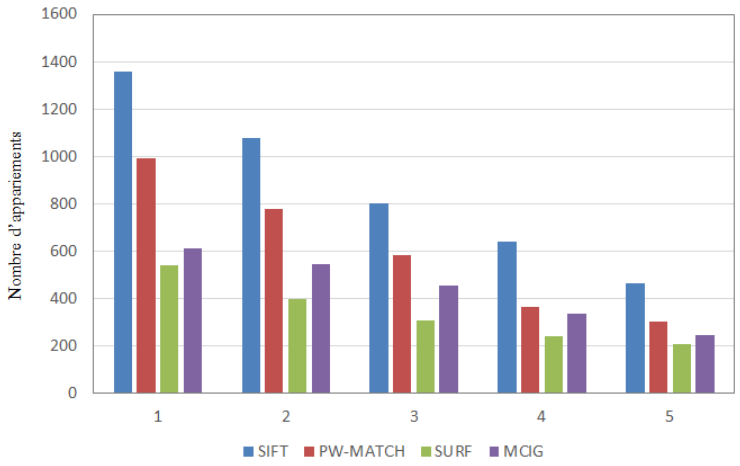

Figure 3.11. Résultats comparatifs en termes de nombre d'appariements trouvés lors des transformations de changement de luminance

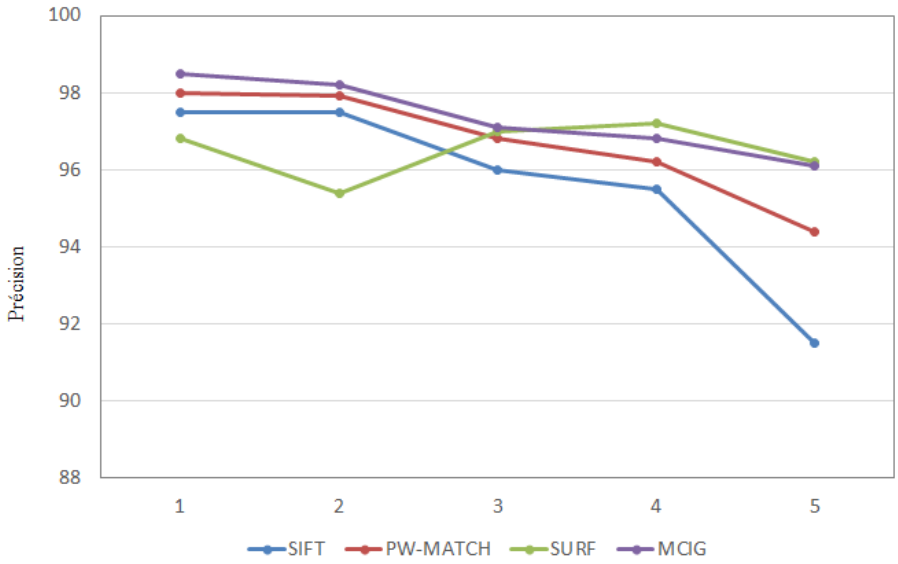

Figure 3.12. Résultats en termes de précision lors des transformations de changement de luminance

D'après la figure 3.12, lors d'un changement de luminance, la méthode proposée MCIG présente un meilleur résultat que le SIFT, SURF et PW-MATCH en termes de précision. Cela est plus visible pour les changements faibles de luminance. Pour les changements de luminance plus importants, le résultat est meilleur que le SIFT et PW-MATCH et il est presque égal au SURF.

En ce qui concerne le nombre d'appariements, il est clair que le nombre d'appariements de la méthode MCIG est meilleur que celui de SURF. Ce nombre est plus faible que le SIFT et PW-MATCH mais l'avantage est qu'il décroit moins rapidement.

3.1.2.3 Estimation du temps d'exécution

Dans le tableau 3.4, nous évaluons la robustesse de la méthode MCIG proposée en terme du temps d'exécution en comparaison avec les méthodes SIFT, SURF et PW-MATCH. Pour ce faire, nous avons pris un couple d'image ayant une transforma-

tion moyenne (image 1 et image 4) de chacune des séquences d'images (Graffiti, Boat, cars) présentées respectivement dans les figures 3.4, 3.7 et 3.10. A partir de la

Méthode	Boat	cars	Graffiti	Moyenne
SIFT	2.5857	2.1471	2.7107	2.4811
SURF	1.9018	2.0251	1.909	1.9453
PW-MATCH	3.2321	2.5367	3.188	2.9856
MCIG	2.2716	2.1105	2.2837	2.2219

Table 3.4. Tableau comparatif des résultats obtenus en termes du temps de calcul (en milliseconde) entre la méthode proposée et les méthodes SIFT, SURF et PW-MATCH

dernière colonne du tableau 3.4, on peut remarquer que les résultats de la méthode proposée MCIG sont nettement meilleur que SIFT et PW-MACTH en termes de temps d'exécution. Ces résultats sont légèrement inférieur à SURF. En conclusion, la méthode de SURF est la plus rapide suivi par la méthode MCIG puis SIFT et à la fin PW-MATCH.

3.1.3 Discussion

On peut noter que la méthode de mise en correspondance MCIG proposée montre un résultat très compétitif. En effet, elle montre un meilleur compromis entre le nombre des appariements, la précision et le temps d'exécution. On remarque que la précision, en comparaison avec les méthodes existantes, est soit égale soit meilleure dans les faibles changements. Cependant, elle est toujours meilleure pour les changements plus importants. En plus la comparaison en terme de temps d'exécution est très satisfaisante. Ceci ne peut que prouver une meilleur stabilité face aux différents changements et par la suite montrer l'apport de la description locale par LBP et

l'appariement par ajout des contraintes géométrique. On peut remarquer aussi que les courbes des précisons des méthodes SIFT décroisent rapidement même si que le nombre des appariements reste important avec un temps d'exécution relativement élevé. Ceci affirme que SIFT en plus qu'il est lent, génère un nombre élevé d'appariements et par conséquent un nombre élevé de faux appariements. Ainsi, la méthode proposée a réussi, d'après les trois métriques utilisées : celle du nombre des appariements, celle de précision et celle du temps d'exécution, à remédier ce problème. Donc, tous les tests confirment le choix de la méthode MCIG pour utiliser dans le reste de nos travaux.

3.2 Extraction des images clés

Dans ce qui suit, nous allons expliquer le protocole d'évaluation que nous allons suivre afin de montrer la qualité des images clés extraites.

3.2.1 Protocole d'évaluation

Nous avons suivi une méthodologie d'évaluation de la robustesse des résumés construis pour chacune des méthodes proposées EICCTR EICGR-1 et EICGR-2. Cette méthodologie se base sur une combinaison de critères subjectifs (la qualité) et objectifs (la quantité). Dans un premier lieu, nous avons établi une évaluation subjective qui consiste à juger si le résumé généré contient des segments importants en comparaison avec le contenu de la vidéo originale. La notion de subjectivité vient en partie d'une comparaison des segments générés automatiquement en appliquant les méthodes proposées et ceux de la vérité terrain fournie par la base. Dans une seconde étape, nous passons à une évaluation objective. Cette dernière consiste à mesurer les performances des systèmes proposés tel que le temps pris pour la génération des résumés, le taux de compression et le rapport signal/bruit. Dans ce contexte, nous allons montrer les résultats obtenus pour quelques vidéos appartenants aux deux bases

utilisées pour tester les résultats. Ainsi, dans la comparaison subjective, nous allons comparer les résultats obtenus avec les résumés fournis par la base OVP ainsi que quatre méthodes importantes dans la littérature qui ont été discutées au préalable dans le chapitre de l'état de l'art et qui ont utilisé la base OVP dans leurs processus expérimentaux :

- DT [Mundur, 2006] : utilise l'algorithme de Triangulation de Delaunay pour classifier les images des vidéos puis le centre de chaque classe sera inséré dans l'ensemble des images clés.

- STIMO (STIll and MOving Video Storyboards) [Furini, 2010] : cette méthode génère un résumé statique en utilisant l'histogramme de couleur HSV. Il utilise l'algorithme de variation moyenne FPF (Farthest Point-First).

- VSUMM (Video summarization) [Sandra, 2011] : algorithme simple d'extraction de résumé basé sur la classification k-moyenne et l'histogramme HSV pour caractériser la couleur.

- VISCOM (Video Summarizationusing Colorco-Occurrence Matrices [Vinicius, 2017] : Une méthode qui utilise les matrices de cooccurrences des couleurs dans le processus de sélection des images clés.

Pour la comparaison objective, sachant que les quatre méthodes utilisées dans la comparaison subjective se basent sur la description globale, nous allons comparer les méthodes proposées avec la méthode (parmi les quatre méthodes de la littérature citées précédemment) qui a donné un meilleur résultat dans la comparaison subjective avec 2 autres méthodes utilisant la description locale pour la génération des résumés statiques:

- Méthode proposée par [Tapu, 2011] : Cette méthode est basée sur la distance de x^2 des histogrammes de couleur HSV et le descripteur SIFT.

- Méthode proposée par [Massaoudi, 2017] : Cette méthode est basée sur la description locale utilisant le détecteur SURF et la méthode FLANN pour la sélection des images clés.

Cette comparaison va nous permettre de mettre en valeur des méthodes proposées (qui sont aussi basées sur la description locale) par rapport à ceux appartenant à la même famille dans la littérature.

3.2.1.1 Base des vidéos

Afin de vérifier leur efficacité, nous avons évalué les méthodes d'extraction d'images clés proposées sur différents types de vidéos (films, journal, cartoons, jeux...). Ces vidéos présentent plusieurs challenges comme le mouvement de la camera, l'arrière-plan dynamique, etc. Nous avons commencé avec des tests d'évaluation qualitative puisque le jugement subjectif est très efficace dans ce cadre et il est très utilisé dans la littérature. Ensuite, nous avons enchainé avec des tests quantitatifs en utilisant le rapport signal bruit et le taux de compression. L'ensemble des vidéos de tests comprend des séquences des deux bases suivantes qui contiennent des vidéos caractérisées par un contenu diversifié (documentaire, pédagogique, conférence, dessins animés et historique). Chaque vidéo appartenant aux deux bases a été divisée au préalable en plans à l'aide de la méthode basée sur la distance chi carré des histogrammes [Cai, 2005].

- "YUV" (YUV Video Sequences - http://trace.eas.asu.edu/yuv/) : Cette base a été choisie pour la richesse des vidéos en termes de résolutions et contenu diversifié. Ces vidéos sont composées d'images de tailles différentes : 352 x 240 ; 176 x 144 et 352 x 288. Le tableau 3.5 présente la durée, le nombre d'images et de plans pour quelques vidéos de la base YUV.

- " OVP" (The Open Video Project (2016) - http://www.open-video.org) : Nous

Titre de la vidéo	Durée(mm:s)	Nombre des images	Nombre de plans
News	0:12	300	2
Foreman	0:12	297	3
Mother and Daughter	0:12	300	1
Filinstone	0:16	510	10
Carphone	0:12	382	2

Table 3.5. Exemples de vidéos de la base YUV et leurs caractéristiques.

avons utilisé un ensemble de 50 vidéos appartenant à cette base conçue pour l'évaluation de résumé. Ces vidéos sont dans le format "MPEG-1" avec 30 images par seconde. Chaque image est de taille 352 x 240 pixels avec une durée qui varie entre une et quatre minutes. Nous avons choisi cette base vu qu'elle fournit au préalable des résumés pour chaque vidéo. Ces résumé sont considérés comme une vérité terrain "OVP summaries". Le tableau de la figure 3.13 présente la durée, le nombre d'images et de plans pour quelques vidéos de la base OVP.

Titre de la vidéo	Genre	Durée (mm:ss)	Nombre d'images	Nombre de plans
The great web of water, segment 01	Documentaire	1:50	3279	31
Sense and sensitivity introduction to lecture 2	Lecture	1:53	3411	7
The future of energy gases, segment 09	Documentaire	1:02	1884	6
America's New Frontier, Segment 10	Documentaire	2:41	4830	9
Digital Jewelry: Wearable Technology for Every Day Life	Educationnel	3:00	4204	10
Exotic Terrane, segment 01	Documentaire	1:38	2940	16

Figure 3.13. Exemples de vidéos de la base OVP et leurs caractéristiques.

3.2.1.2 Métriques utilisées

Nous avons utilisé une sélection de métriques proposées dans le protocole d'extraction des résumés des vidéos statiques (images clés) [Sandra, 2011]. Ces métriques permettent l'évaluation de l'efficacité et la faisabilité des résultats obtenus par rapport aux vidéos d'origine ainsi que la performance globale du résumé généré. Ces résultats seront comparés par rapport à des méthodes récentes de l'état de l'art ainsi qu'avec la vérité terrain fournie par la base OVP. Pour cela, nous détaillons dans ce qui suit les métriques utiles à ce protocole.

- Comparaison aux résumés d'utilisateurs (CUS) métrique [Sandra, 2011]

Dans cette méthodologie proposée par Sandra et al., [Sandra, 2011] puis améliorée par Vinicius et al., [Vinicius , 2017], le résumé automatique produit sera comparé avec des résumés construits par des utilisateurs (ensemble de 5 résumés produits manuellement par 5 différents utilisateurs). Les résumés des utilisateurs, dans cette méthodologie, sont considérés comme des références : vérité terrain. La figure 3.14 montre un exemple illustratif des résumés produits par 5 utilisateurs pour la vidéo "The Future of Energy Gases, segment 09 (v53)". Si deux images étaient considérées similaires, la première image appartenant au résumé statique et l'autre au résumé utilisateur, alors elles seraient supprimées des prochaines itérations de CUS. Le concept de similarité est basé sur le calcul de la distance de Manhattan entre les histogrammes couleurs des images [Swain, 1991]. En effet, l'histogramme de couleur est généralement appliqué pour décrire le contenu visuel des images vu sa complexité de calcul triviale. Il est important à noter que deux images clés ne doivent pas être assez identiques pour être considérées assez similaires. Ainsi, le seuil a été fixé à une valeur égale à 0,5. (Si le seuil est inférieur à 0,5 les deux images clés sont considérées similaires) [Sandra, 2011]

L'évaluation sera basée sur trois paramètres :

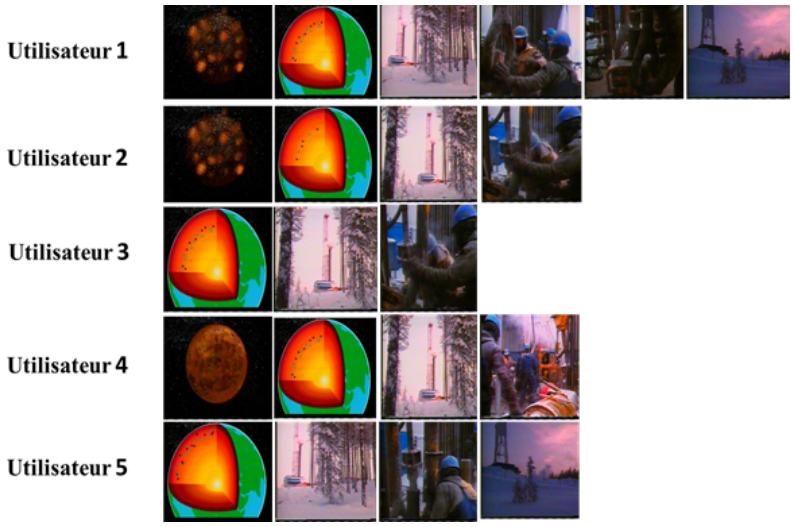

Figure 3.14. Résumés des 5 utilisateurs pour la séquence de vidéo "The Future of Energy Gases, segment 09 (v53)".

- SF_i: est le nombre des images similaires correspondantes entre le résumé automatique et le résumé utilisateur.
- AS_i : le nombre des images produites par le résumé automatique.
- US_i : Le nombre des images existantes dans les résumés des utilisateurs.
  avec $i \in 1, 2, 3, 4, 5$ relative à un utilisateur spécifique

La valeur de précision $P_i = SF_i/AS_i$ et celle du rappel $R_i = SF_i/US_i$ seront calculés à l'aide des paramètres cités précédemment. Par la suite, la métrique "F-measure" sera calculée pour chaque vidéo. Cette mesure représente la moyenne harmonique entre la précision et le rappel (les valeurs de P_i etR_i), comme indiqué dans l'équation 3.1 :

$$F - measure = \frac{\sum_{i=1}^{5} \frac{2 \times P_i \times R_i}{P_i + R_i}}{5} \tag{3.1}$$

Dans la figure 3.15, on montre un exemple illustratif du processus de la comparaison du résumé généré automatiquement avec celui des utilisateurs.

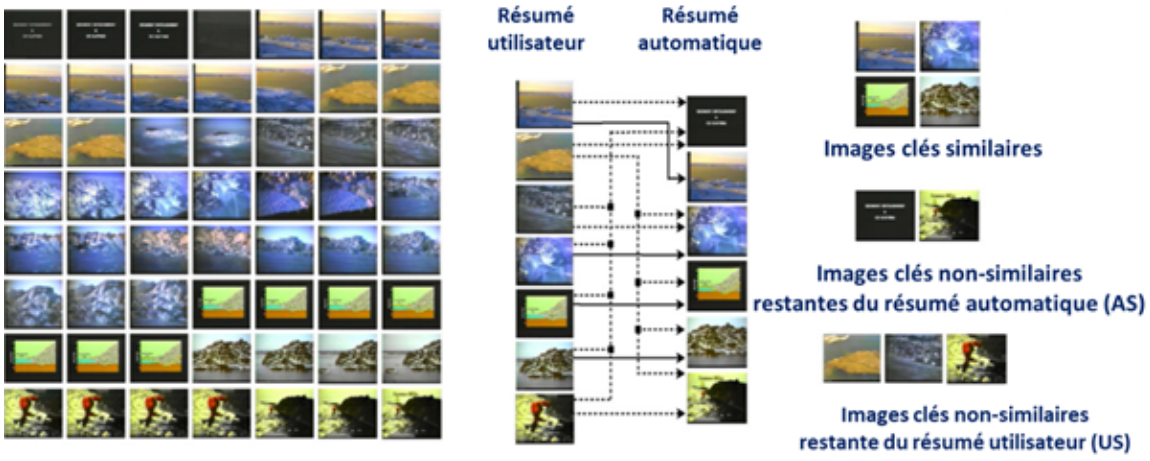

Figure 3.15. Exemple illustratif de la méthodologie d'évaluation CUS (Comparaison aux résumés d'utilisateurs)

- Taux de compression (CR%)

Le résultat de l'extraction d'images clés doit être compact afin d'éviter la redondance. Dans ce contexte, nous avons utilisé le taux de compression pour vérifier ce critère. Ce taux CR% est calculé par la division du nombre d'images clés par le nombre d'images de la vidéo. Pour une séquence donnée, le taux de compression est défini par l'équation 3.2 :

$$CR = 1 - \frac{card\{keyframes\}}{card\{frames\}} \tag{3.2}$$

où $card\{keyframes\}$ est le nombre d'images clés extraites de la vidéo et $card\{frames\}$ est le nombre d'images de la vidéo.

- Rapport signal/bruit ((RSB) ou (PSNR))

Nous avons calculé le RSB pour chaque couple (F_u, F_v) d'images clés de taille (M*N) extraites. Ensuite, nous considérons la moyenne des RSBs pour chaque vidéo.

$$RSB(F_u, F_v) = 10\,log - \left(\frac{(N * M * 255)^2}{\sum_{x=1}^{N} \sum_{y=1}^{M} (F_u(x,y) - F_v(x,y))} \right) \tag{3.3}$$

Plus que les images clés F_u et F_v sont similaires, plus la valeur PSNR est élevée. Les valeurs infinies du PSNR reflètent une redondance des images clés extraites et les valeurs réduites indiquent leur diversité.

3.2.2 Évaluation qualitative

Il est important de pouvoir évaluer la qualité des résumés générés automatiquement. Cependant, l'évaluation de la qualité des résumés générés est une tâche délicate. C'est l'une des parties les plus difficiles à mettre en place dans le processus de développement de méthodes de création des résumés vidéos. Il est extrêmement difficile de donner une définition formelle de ce qui est un bon résumé. Dans un premier lieu, nous allons montrer quelques exemples de résultats des méthodes d'extraction des images clé EICCTR, EICGR-1 et EICGR-2 proposées, et ce pour les bases "YUV" et "OVP".
Vu que cette dernière dispose d'une vérité terrain, ces résultats seront suivis par différentes mesures de métriques utilisés tels que le rappel, la précision et la F1-mesure, ceci pour les résumés trouvés automatiquement à travers les différentes méthodes proposées en comparaison avec d'autres méthodes de l'état de l'art. Nous allons monter par la suite pour quelques vidéos appartenant aux deux bases de tests les résultats en termes de temps d'exécution normalisé (en seconde). Tous ces résultats nous permettrons d'établir une étude comparative subjective entre les différentes méthodes proposées.
Les figures 3.16, 3.17 et 3.18 montrent des exemples des résultats des images clés pour quelques vidéos appartenant aux deux bases.

Nous présentons dans la figure 3.16 les images clés de la vidéo "finlinstone" appartenant à la base "YUV". Cette vidéo contient 510 images. En appliquant la méthode proposée EICCTR nous réussissons à extraire 14 images clés, 13 images clés sont extraites en appliquant la méthode EICGR-1 et 12 images clé résultent lors de l'application de la méthode proposée EICGR-2.

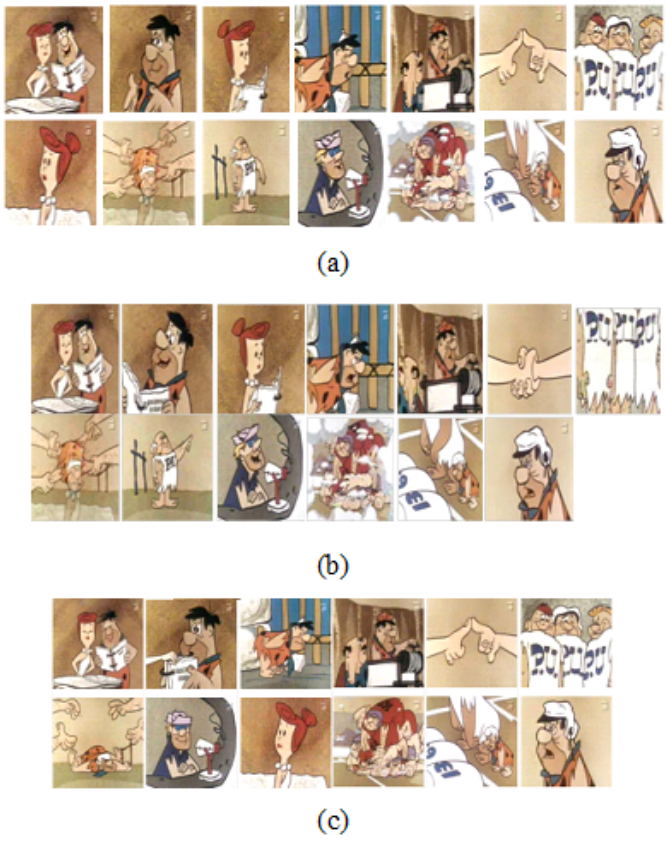

Figure 3.16. Images clé produites par les différentes méthodes proposées pour la vidéo "Filinstone" tel que (a), (b) et (c) sont respectivement relatives aux méthodes EICCTR,EICGR-1 et EICGR-2

Dans la figure 3.17, nous montrons les images clés obtenues par chacune des méthodes proposées de la vidéo “Foreman” appartenant aussi à la base "YUV". Cette vidéo aura : 5, 4 et 3 images clés en appliquant respectivement les méthodes proposée EICCTR, EICGR-1 et EICGR-2. Enfin, dans la figure 3.18, nous montrons les images clés obtenues par chacune des méthodes proposées d'extraction d'images clés de la vidéo "The Future of Energy Gases, segment 09" appartenant à la base "OVP". Cette vidéo aura : 7, 6 et 4 images clés en appliquant respectivement les méthodes proposée EICCTR, EICGR-1 et EICGR-2.

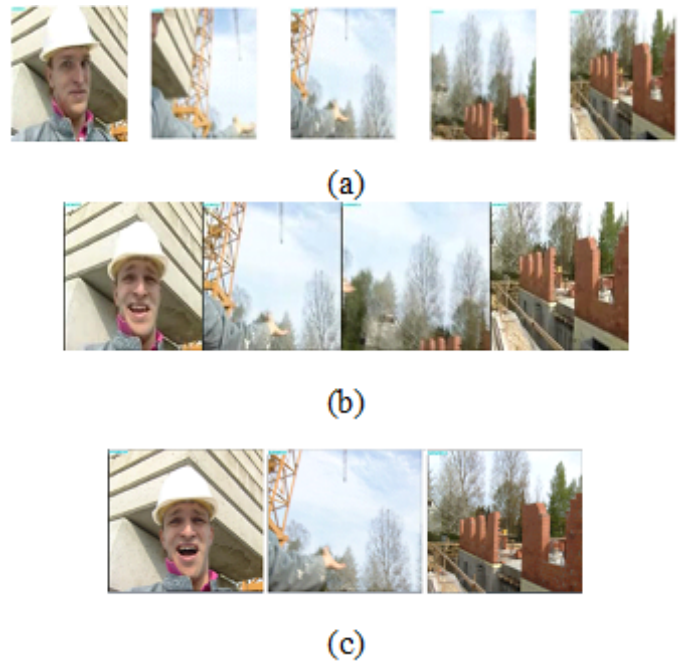

Figure 3.17. Images clés produites pour la vidéo "Foreman.mp4" tel par les différentes méthodes proposées que (a), (b) et (c) sont respectivement relatives aux méthodes EICCTR, EICGR-1 et EICGR-2

Le tableau 3.6 montre les différentes moyennes en termes de précision, rappel pour quelques vidéos choisies pour test de la base "OVP", et ceux pour les méthodes proposées EICCTR, EICGR-1 et EICGR-2 en comparaison avec d'autres méthodes de la littérature. Ce tableau sera suivi d'un diagramme (figure 3.19) qui les résume en termes de F1-mesure. Les résultats présentés dans le tableau 3.6 et la figure 3.19 vont confirmer les bons résultats trouvés dans les figures précédentes des résumés générés.

Comme première lecture du tableau 3.6 ainsi que la figure 3.19, on peut remarquer que les valeurs obtenues pour les méthodes proposées pour l'extraction des images clés, sont en générale bonnes en comparaison avec le reste des valeurs résultantes des autres méthodes existantes dans la littérature. En effet, il est très clair que les résultats des méthodes proposées EICCTR, EICGR-2 surmontent les méthodes exis-

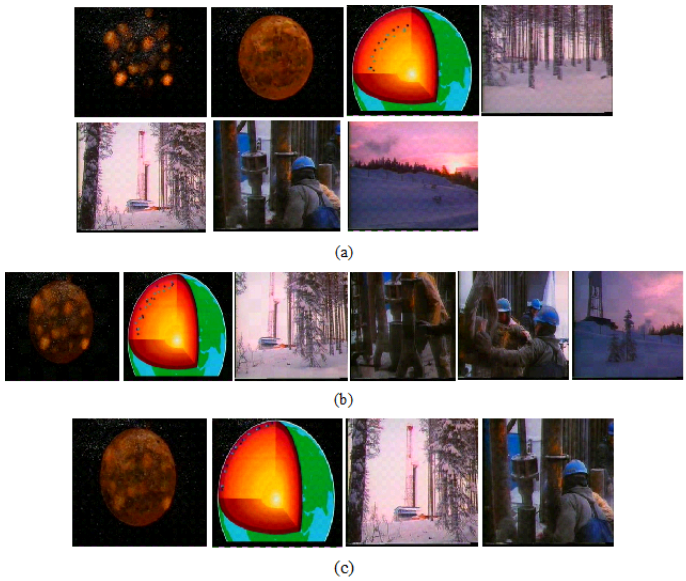

Figure 3.18. Images clé de la vidéo "The Future of Energy Gases, segment 09" produites par les différentes méthodes proposées tel que (a), (b) et (c) sont respectivement relatives aux méthodes EICCTR, EICGR-1 et EICGR-2

tantes en termes de F1-mesure. Les résultats de la méthode EICGR-1 sont bonnes aussi par rapport à la littérature (meilleures que DT et STIMO et presque similaires à OVP et VSUMM) mais légèrement inférieures à la méthode existante VISCOM et aux deux autre méthodes proposées EICCTR, EICGR-2.

En ce qui concerne le temps de calcul, la figure 3.20 montre que la méthode EICGR-2 est la moins couteuse en termes de temps d'exécution. En second lieu, arrive la méthode EICGR-1 puis EICCTR. Ainsi, on peut noter que cela est dû au fait que le traitement dans la méthode EICCTR s'effectue sur toute les images de la vidéo en plus du passage par le découpage de la vidéo en plans et le passage par la réduction de dimension. De même pour la méthode EICGR-1 qui vient en deuxième place, cela peut être dû aux passage par l'étape de segmentation de la vidéo en plans en plus de celle de la recherche de la valeur minimal pour chaque table relative aux différents plans. Les résultats de la méthode EICGR-2 prouvent davantage que le traitement de

	Précision (%)	Rappel (%)
OVP	58.4	65.7
VSUMM	72.1	64.1
DT	54.7	43.3
STIMO	51.9	62.1
VISCOM	64,9	81,1
EICCTR	68.6	72,1
EICGR-1	62.7	58.9
EICGR-2	66.1	79.8

Table 3.6. Valeurs moyenne en termes de précision et rappel des images clés produites, pour chacune des méthodes, pour toutes les vidéos choisies pour test de la base "OVP"

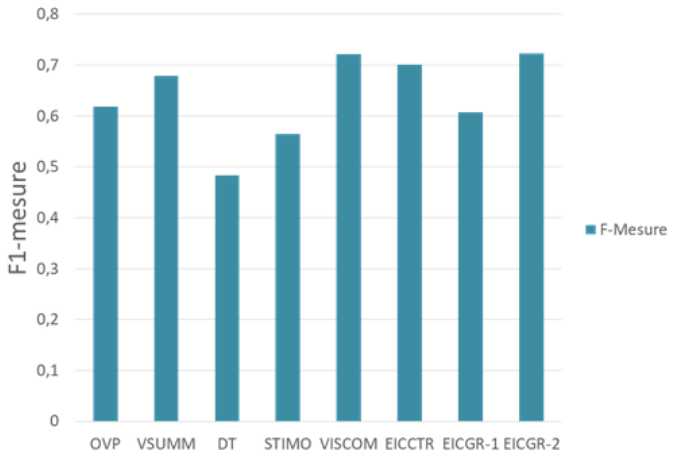

Figure 3.19. Moyenne des valeurs de F-mesure des images clés produites, pour chacune des méthodes, pour toutes les vidéos choisies pour test de la base "OVP"

toute la vidéo sans passage par décomposition par plans et par la suite la génération des images candidates pour toute la vidéo, en plus du la classification utilisant le critère de modularité ont minimisé considérablement le temps de calcul sans affecter la qualité des résultats.

En conclusion, d'après les deux figures 3.19 et 3.20, il est clair que la méthode

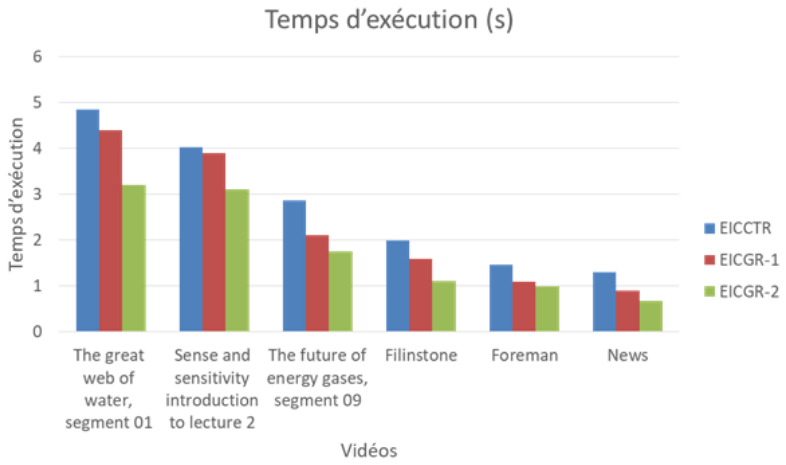

Figure 3.20. Résultats en termes de temps d'exécution de quelques vidéos choisies pour test des deux bases "OVP" et "YUV" et ce pour méthodes proposée EICCTR, EICGR-1 et EICGR-2

proposée EICGR-2 présente un compromis de qualité entre la mesure F1 et le temps d'exécution (aussi bien de complexité). La méthode EICCTR donne aussi des bons résultats mais avec un cout un peu plus élevé en termes de temps d'exécution et de complexité. Pour remédier à ce problème, nous avons proposé EICGR-1 qui a minimisé ces deux contraintes mais nous avons perdu en termes de qualité. Ainsi, EICGR-2 a été proposée pour améliorer à la fois le temps de calcul, la complexité ainsi que la qualité.

3.2.3 Évaluation quantitative

Afin d'évaluer quantitativement les résultats des résumés statiques produits par les trois variantes des méthodes proposées, nous déterminons les métriques d'évaluation de taux de compression (CR%) et de rapport signal bruit (PSNR) citées déjà dans la sous-section 3.2.1. Nous comparons ainsi les valeurs obtenues avec trois autres méthodes : celle fournie par la méthode de VISCOM [Vinicius, 2017] (qui contribue avec des résultats importants dans le domaine d'extraction d'images clés et qui a donné un meilleur résultat en termes de qualité par rapport aux différentes méthodes testées dans la littérature) et celles proposées par Tapu et al., [Tapu, 2011] et Massaoudi et al.,[Massaoudi, 2017] qui sont basées sur la description locale. Les résultats obtenus sont reportés dans les figures 3.21 et 3.22.

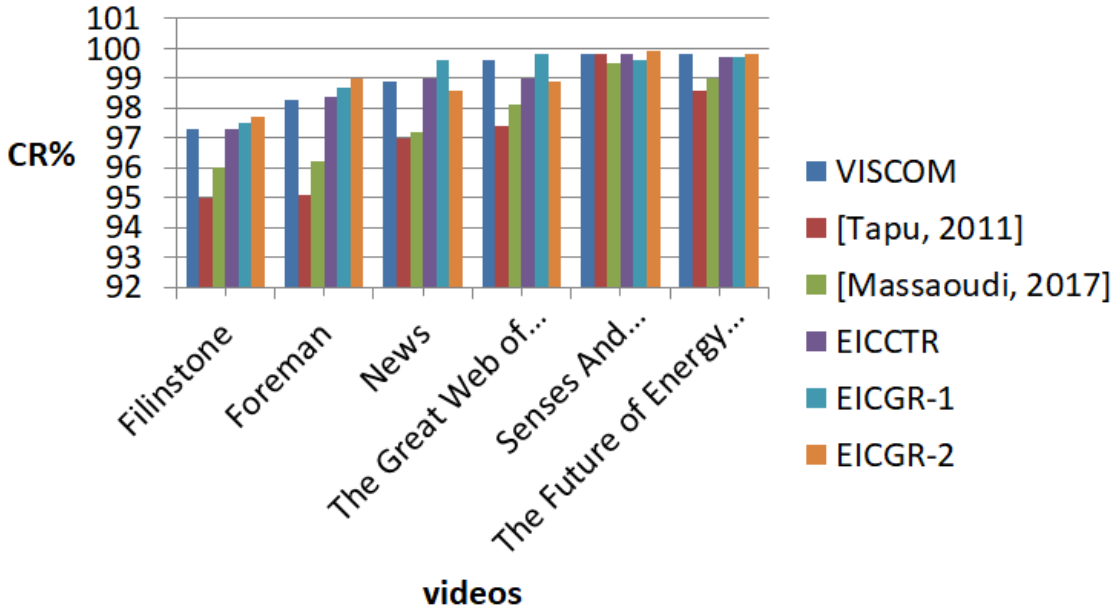

Figure 3.21. Comparaison de la qualité des résultats obtenus en termes de taux de compression

A partir de la figure 3.21, nous constatons que les méthodes proposées nous four-

nissent des taux de compression élevés (toujours supérieurs à 97.3%). Ces taux sont pour la majorité de vidéos supérieures (parfois égaux) au taux trouvé pour VISCOM qui se base essentiellement sur la description globale. Cependant, ils sont largement supérieurs aux méthodes proposées par Tapu et al. et Massaoudi et al. qui sont basées sur la description locale. En effet, une valeur du CR% élevée indique que les images clés produites sont différentes et par la suite une réduction considérable dans la redondance des images clés extraites.

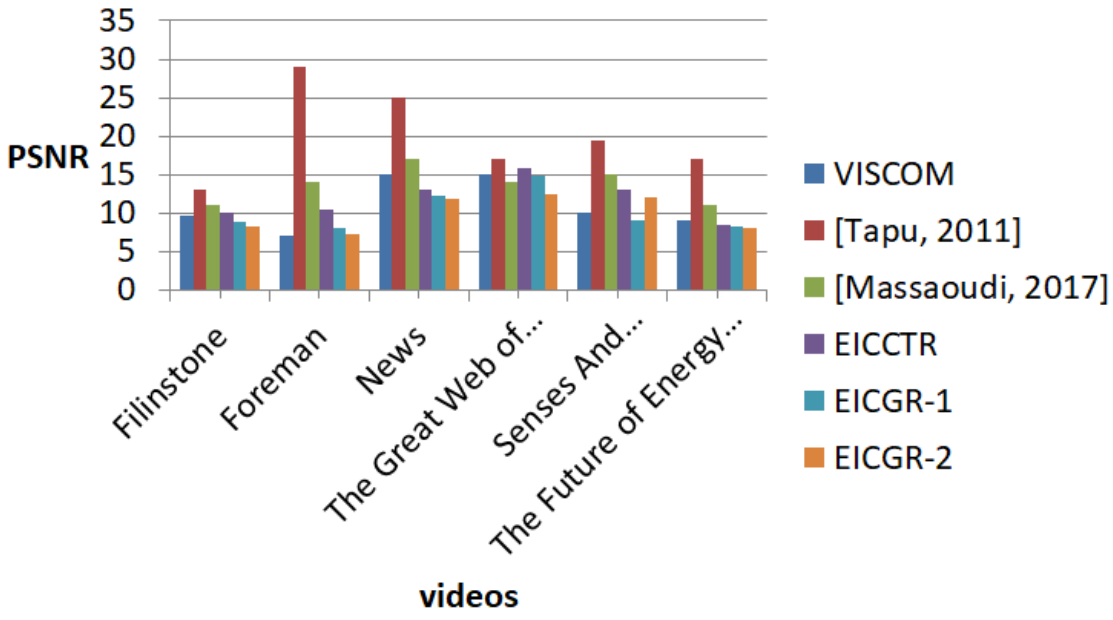

Figure 3.22. Comparaison de la qualité des résultats obtenus en termes de taux de PSNR (Rapport signal sur bruit)

Ces résultats confirment davantage notre supposition initiale qui considère que la description locale par points d'intérêts est une bonne solution pour l'extraction des images clés, vu la robustesse du processus d'extraction des points d'intérêts face à différentes transformations.

Dans la figure 3.22, on montre une comparaison des méthodes proposées avec celles

de la littérature en termes de PSNR. Une première lecture du graphique, permet de remarquer la réussite du processus d'extraction. En effet, les images clés sont considérées similaires, si la valeur du PSNR est élevée. Donc, des valeurs élevées de PSNR reflètent une redondance entre les images clés extraites. Inversement, des valeurs réduites indiquent une diversité de ces images clés en termes de contenu. Il est clair que les valeurs enregistrées de PSNR sont faibles, ceci est confirmé en les comparants avec les valeurs relatives aux autres méthodes. Ces résultats confirment que notre méthode extrait les images clés les plus significatives et pertinentes ce qui favorise la minimisation de la redondance.

3.2.4 Discussion

Dans la section précédente, nous avons traité la problématique du résumé vidéo statique. En effet, le résumé de vidéo sous la forme d'un ensemble d'images clés est une étape essentielle qui facilite le processus de recherche de vidéo par le contenu. Dans ce contexte, nous avons proposé trois méthodes de résumé. Il est clair d'après la comparaison des méthodes proposées qu'elles surpassent des méthodes qui contribuent bien dans la littérature. Cette comparaison les a bien mises en valeur. Ceci est valable pour l'évaluation subjective aussi bien que pour l'évaluation objective. Ces résultats sont obtenus grâce aux avantages de la description locale par points d'intérêt et par la suite à la méthode de mise en correspondance MCIG proposée. Chacune de ces variantes possède des avantages et des inconvénients. D'après l'évaluation subjective, on peut dire que la méthode EICGR-2, prend les avantages du traitement des images candidats de la vidéo entière et de l'utilisation de la classification de graphe par maximisation de modularité pour donner un compromis entre la qualité des résultats et le temps d'exécution. Ainsi, nous avons atteint notre objectif qui consiste à définir un ensemble d'images significatives qui sont considérées comme représentatives par rapport au contenu en informations d'une séquence vidéo donnée en se basant sur la description locale par points d'intérêts.

3.3 Prototype proposé d'un système de recherche de vidéos par le contenu

Pour atteindre notre objectif initial, qui consiste à exploiter la description locale pour la génération d'un résumé statique et faciliter le processus de recherche de vidéos par le contenu, il est très important de projeter nos résultats sur les perspectives du domaine de recherche de vidéos par le contenu pour s'assurer davantage de l'efficacité de la méthodologie proposée. En effet, il est intéressant de savoir quelles sont les méthodologies qui permettraient d'améliorer la qualité de recherche et de se rapprocher aux maximum des exigences de l'utilisateur. Vu que la méthode EICGR-2 proposée donne un meilleur compromis entre précision, temps d'exécution et représentativité, elle sera choisie pour être testée dans un prototype de système de recherche de vidéos par le contenu.

3.3.1 Architecture générale du prototype

Deux étapes indissociables coexistent dans le système de recherche de vidéo par le contenu. Comme première étape, les vidéos de la base seront décrites par leurs images clés. Ainsi, nous obtenons la base des images clés. Nous passons par la suite à l'étape de description locale pour les différentes images clés de chaque vidéo. Cette étape se base essentiellement sur la détection de points d'intérêts à l'aide de détecteur SIFT. Puis l'extraction des descripteurs locaux autour des images clés détectées. Ces étapes sont traitées hors ligne. Par la suite, dans la partie en ligne, lorsque l'utilisateur entre une image requête. Les descripteurs locaux de cette image seront calculés. Ainsi, afin d'obtenir les vidéos résultats, un calcul de similarité entre l'image requête et les images clés de chaque vidéo sera effectué et le système donne la liste des vidéos résultats en utilisant une mesure de similarité. Cette mesure sera décrite dans le paragraphe suivant . La figure 3.23 montre le schéma de test ainsi que les différentes étapes de recherche.

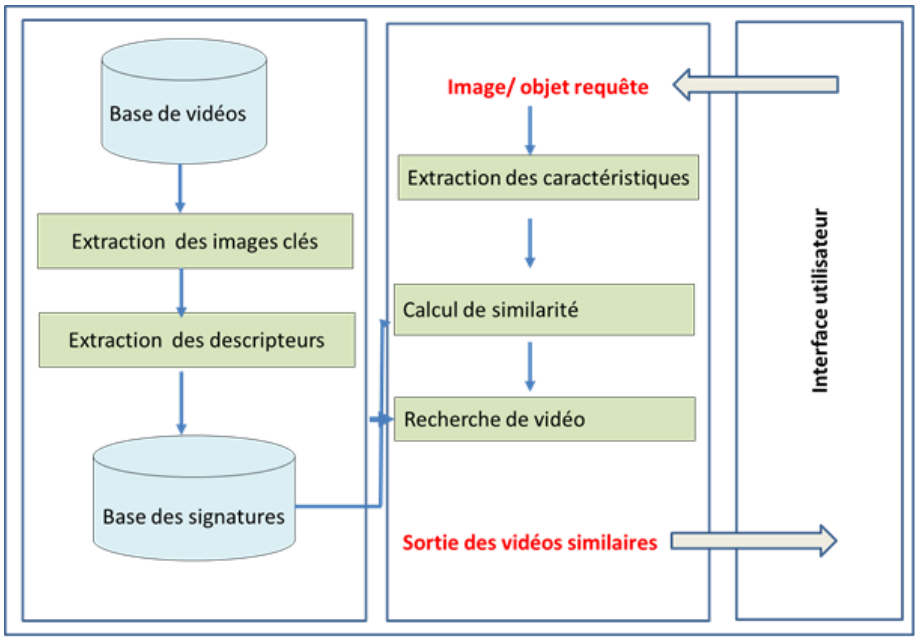

Figure 3.23. Schéma illustratif de l'évaluation des deux méthodes proposées dans le contexte de la recherche par le contenu

3.3.2 Mesure de similarité proposée

La mesure de similarité adoptée c'est la mesure de répétabilité (la même utilisée dans le processus d'extraction des images clés). Ainsi, le principe consiste à calculer la répétabilité maximum entre l'image requête et les différentes images clés de chaque vidéo. Si cette valeur maximale est supérieure à un certain seuil alors la vidéo sera affichée avec l'ensemble des résultats. Ainsi, le seuil S a été fixé pour une valeur de 50 % après plusieurs expérimentations. Ce choix de la valeur du seuil n'a pas été strict parce que l'utilisateur n'aura pas besoin seulement des vidéos contenant un contenu complètement similaire à sa requête mais aussi une petite similarité pourrait lui intéresser. Nous montrons dans la figure 3.24, un exemple des valeurs de répétabilité entre une image requête et des images appartenant à la base utilisée.

Nous pouvons ainsi remarquer, que les images ayant une valeur de répétabilité inférieure

à 50 % sont assez différentes en comparaison avec l'image requête. Tans dis que celle ayant une valeur supérieure à 50 % contiennent des zones de similarité qui peuvent intéresser l'utilisateur.

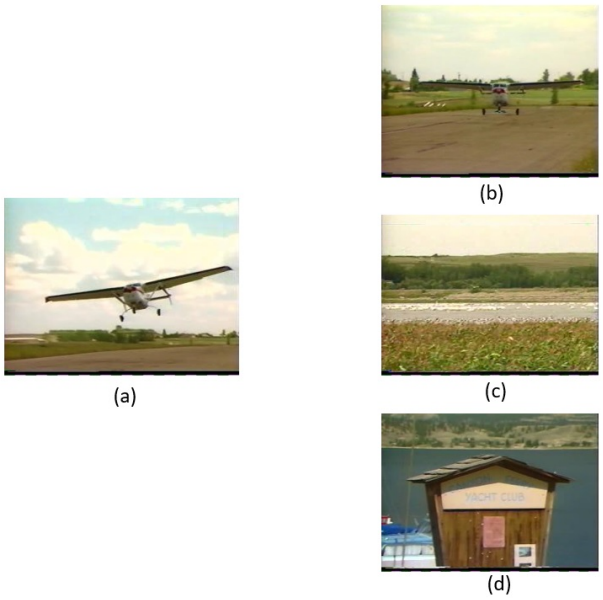

Figure 3.24. image référence (a) et les valeurs de répétabilités avec les images (b), (c) et (d) respectivement 0.54 , 0.2 et 0.01

3.3.3 Protocole d'évaluation

Les deux mesures : rappel et précision sont les métriques les plus communes employées pour mesurer l'efficacité des systèmes de recherche par le contenu. Ils sont basés sur la notion d'ensemble. La mesure rappel montre la capacité de pouvoir récupérer toutes les vidéos recherchées à partir d'une requête par un système. La précision montre la capacité du système à afficher seulement les vidéos appropriées. Ainsi, chaque requête peut être associée à une valeur de précision et à une valeur de rappel sur une

collection donnée de vidéos.

$$Rappel = \frac{card(\{Vidéospertinentes\} \cap \{Vidéosretrouvées\})}{card(\{Vidéospertinentes\})} \tag{3.4}$$

$$Précision = \frac{card(\{Vidéospertinentes\} \cap \{Vidéosretrouvées\})}{card(\{Vidéosretrouvées\})} \tag{3.5}$$

Pour faire une évaluation objective, nous proposons aussi de calculer aussi la courbe de Précision/Rappel. En effet, cette courbe nous permet de suivre la qualité du résultat obtenu en fonction du nombre des vidéos retournés.

Pour tester davantage l'efficacité des résumés vidéo produits par la méthode EICGR-2, nous avons effectué nos tests sur un ensemble de vidéos appartenant à la base OVP . Cette base, vu la diversité du contenu de ses vidéos, est destinée pour l'évaluation de divers processus y compris la recherche par le contenu. Pour la base d'apprentissage, nous avons choisi un ensemble 50 images parmis les images constituant la base de vidéo. L'ensemble des vidéos résultantes pour chacune de images entrées comme requête est connu à priori. Pour la base de test, nous avons utilisé l'ensemble des vidéos appartenant à OVP (80 vidéos) utilisés par Trecvid (http:// www-nlpir.nist.gov/ projects/trecvid/collection.html) dans plusieurs type de processus y compris celui de recherche.

3.3.4 Évaluation quantitative

Dans le contexte de l'évaluation subjective, nous allons présenter dans les figures 3.25, 3.26, 3.27 et 3.28 des exemples des entrées (images requêtes) et des résultats des vidéos obtenus.

3.3.5 Évaluation qualitative

Pour faire une évaluation objective, nous nous baserons sur la courbe de Précision/Rappel. Les valeurs de Rappel/Précision peuvent être calculées pour chaque requête à part. Mais dans le but de stabiliser l'exécution d'un système de recherche, l'évaluation doit

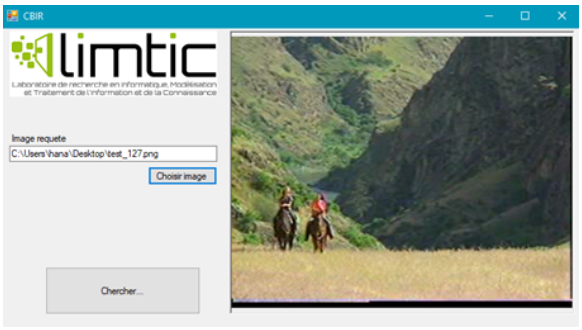

Figure 3.25. Exemple d'image entrée comme requête

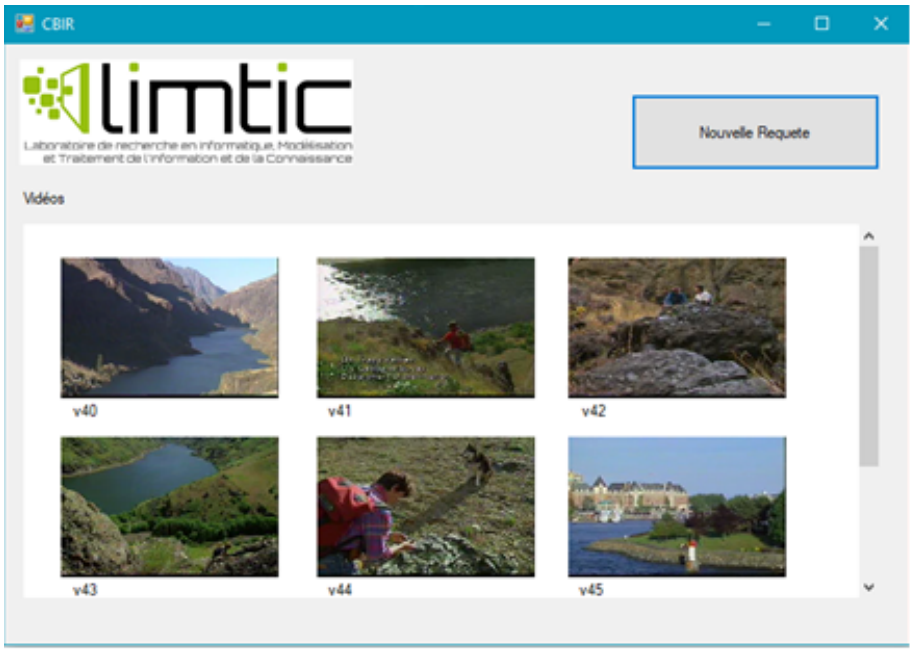

Figure 3.26. Résultats de recherche : les six premières vidéos obtenues lors de l'entrée de l'image requête de la figure 3.25

être faite sur un nombre assez élevé de requêtes. Les performances des systèmes sont alors rapportées sous forme d'une courbe de Précision/Rappel. Ainsi, il est nécessaire d'avoir une compensation entre la mesure de précision et celle du rappel pour une requête donnée : une augmentation dans l'une mène à une baisse dans l'autre. Par

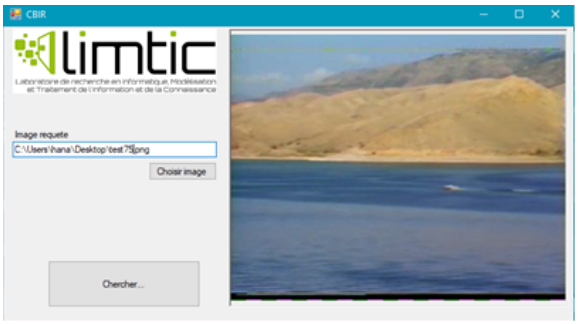

Figure 3.27. Exemple d'image entrée comme requête

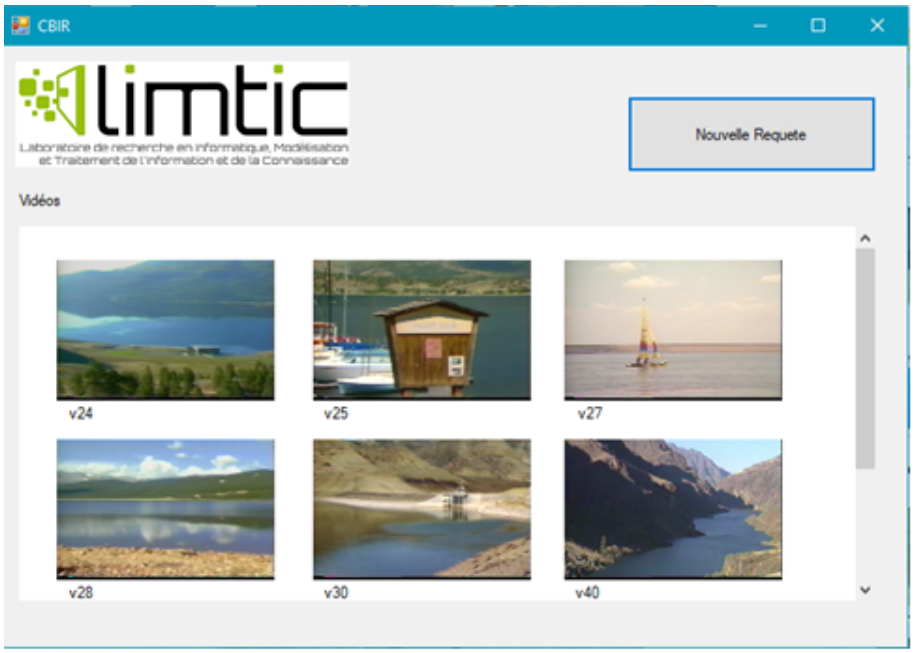

Figure 3.28. Résultats de recherche : les six premières vidéos obtenues lors de l'entrée de l'image requête de la figure 3.27

conséquent, la courbe de Précision/Rappel diminue généralement d'une façon monotone.

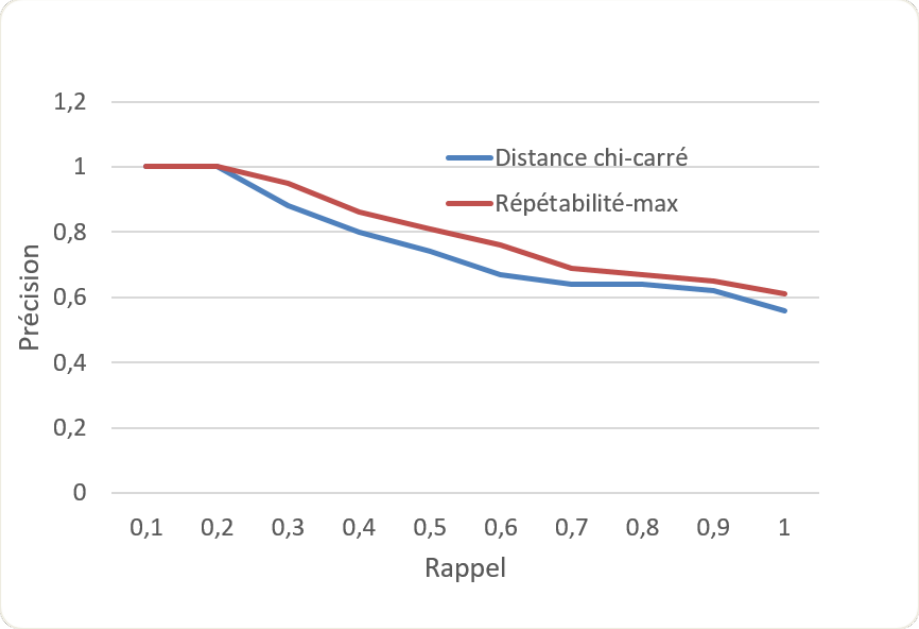

Figure 3.29. Courbe de rappel/précision comparant la recherche en utilisant la mesure de similarité basée sur la répétabilité maximum proposée et celle basée sur la distance Chi-carré des histogrammes.

3.3.6 Discussion

La figure 3.29 montre la courbe de rappel/ précision que nous avons obtenue après une étude expérimentale du système de recherche de vidéos par le contenu. En effet, cette courbe montre l'allure des courbes rappel/précision lors de la recherche en utilisant : la distance basée sur la répétabilité max entre l'image requête et les images clés de la vidéo générés utilisant la méthode EICGR-2 et la distance basée sur la description globale classique utilisant la distance Chi-carré entre l'image requête et les images clés de la vidéo générés utilisant VUSUMM [Sandra, 2011]. Ceci est pour montrer davantage l'efficacité du prototype proposé qui se base essentiellement sur la description local par points d'intérêts. La comparaison des deux courbes se base sur la précision moyenne de chacune. Ainsi, on peut noter que la description locale

par points d'intérêts couplée avec la mesure de similarité que nous avons proposée a réussi à améliorer les résultats de recherche.

Conclusion

Dans ce dernier chapitre, nous avons présenté les différentes expérimentations appliquées dans le cadre de notre travail de thèse. Ainsi, nous avons évalué l'ensemble des résultats des méthodes proposées à savoir la mise en correspondance, l'extraction des images clés. Nous avons défini ainsi un protocole d'évaluation pour chacune des méthodes à part, puis nous les avons testées au sein d'un système de recherche de vidéo par le contenu. Ceci en introduisant une image requête puis essayer de récupérer les vidéos qui ont un contenu similaire. L'objectif principal des méthodes proposée dans ce travail , aussi bien pour la mise en correspondance que pour l'extraction des images clés, est qu'elles soient appliquées sur des bases généralistes et non à des bases spécifiques comme par exemple les vidéos de Sport. Après cette évaluation approfondie des méthodes proposées, il est clair qu'ils ont donné satisfaction dans la base de recherche et ce quelque soit le contenu de la vidéo. Ainsi, nous avons démontré dans ce chapitre que nous avons dépassé l'objectif initial qui consiste à tirer profit des avantages de la description locale par points d'intérêts dans la construction des résumés des vidéos pour faciliter le processus de recherche par le contenu.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Le processus général d'un système de recherche de vidéos par le contenu comprend les étapes suivantes : la description des vidéos, l'indexation et la recherche. Les principaux défis à prendre en considération dans la recherche par le contenu sont liés essentiellement aux caractéristiques utilisées pour la phase de description des vidéos et pour la mesure de similarité.

Dans nos travaux de thèse, nous nous sommes focalisés sur le problème de construction de résumé vidéo. En effet, les résumés vidéo sont considérés parmi les techniques importantes qui permettent la description des vidéos afin de faciliter le processus de recherche par le contenu. Ainsi, le résumé doit répondre à certaines exigences tel que la précision, la fidélité au vidéos originales et la compacité. Pour ce faire, on a besoin des primitives afin d'extraire certaines caractéristiques nécessaires. Etant donné que la description globale a été largement utilisée dans la littérature. Nous avons tiré profit des avantages de la description locale et plus précisément celle basée sur les points d'intérêts. Ainsi, cette description va nous servir à la fois pour la construction des résumés des vidéos en générant les images clés décrivant le contenu visuel et pour la mesure de similarité dans la phase de recherche.

Dans ce mémoire, après avoir présenté l'architecture générale des systèmes de recherche de vidéos par le contenu, nous avons présenté l'état de l'art sur les méthodes de construction des résumés ainsi que les méthodes d'extraction de caractéristiques locales. Pour la description locale, nous nous sommes focalisés des méthodes basées sur les points d'intérêts. Ainsi, nous avons proposé une première méthode qui permet la mise en correspondance des points d'intérêts détectés qui se base sur la description locale par LBP et sur des contraintes spatiales issues des invariants géométriques.

Dans un deuxième lieu, nous nous sommes intéressés à la phase de construction des

résumés statiques des vidéos. La difficulté de l'extraction des images clés est la présence de certaines contraintes telles que le mouvement de caméra, les conditions d'éclairage, etc. Ainsi, nous prenons avantages de la robustesse de la description locale par points d'intérêts face à différentes transformations. Le défi était de monter son efficacité comparée aux autres méthodes surtout celles basées sur la description globale largement utilisée. Une première méthode pour l'extraction des images clés des vidéos a été proposée, cette méthode est basée sur la mesure de répétabilité, couramment utilisée dans la littérature pour juger la ressemblence entre deux images. Une table contenant les valeurs de répétabilités entre chaque deux images pour chaque plan a été construite. La classification de cette table nous a permis de sélectionner les images clés appartenant à chaque plan puis pour toute la vidéo. Une étude expérimentale a été menée sur différentes séquences vidéos de la base de test et a montré une bonne qualité et quantité des images clés produites en comparaison avec les méthodes de l'état de l'art. Néanmoins, cette méthode est coûteuse en termes de complexité.

Dans le but d'améliorer la complexité, on a proposé une deuxième méthode où le traitement s'effectue seulement sur un ensemble d'images sélectionnées par la technique de fenêtrage et non pas sur la totalité des images de la vidéo. Nous avons introduit la notion de graphe pour faciliter la sélection des images clés. Au début la sélection des images clés a été basée sur la répétabilité minimale. Ainsi, les expérimentations ont monté que cette méthode a améloiré les résultats en terme de temps de calcul et de complexité avec une dégradation de la qualité. Puis, nous avons tenté d'améliorer cette méthode pour donner un résultat qui montre un compromis entre la robustesse en termes d'évaluation qualitative, quantitative et complexité. Nous avons sélectionné les images clés en utilisant le principe de classification de graphe par maximisation de la valeur de modularité.

Dans le but de prouver l'efficacité des méthodes déjà proposées vis à vis les méthodes de l'état de l'art, des expérimentations ont été menées sur diverses séquences vidéo. Les résultats obtenus prouvent une meilleure efficacité de la dernière méthode pro-

posée et sa capacité à extraire les images clés les plus représentatives de la vidéo avec un cout raisonnable. Ainsi, pour montrer d'avantage l'efficacité de cette dernière méthode, nous l'avons testée dans un système de recherche de vidéos par le contenu. Les résultats confirment que notre objectif est atteint.

Les travaux menés dans cette thèse évoquent plusieurs perspectives qui peuvent être envisagées. Nous pouvons citer:

- Perspective 1

L'objectif de ce travail étant de fournir un point de départ à l'utilisateur pour initier sa requête dans une grande base de vidéos. L'utilisateur sera perdu dans une grande quantité d'images clés. Il aura du mal à exprimer sa requête. Nous allons essayer dans des travaux futurs de fournir à l'utilisateur un thesaurus visuel composé d'un résumé de l'ensemble des objets les plus répandus dans notre base de vidéo. L'utilisateur pourra ainsi composer une image mentale par une combinaison logique de l'ensemble des objets se trouvant dans le thesaurus visuel. Cette image mentale représentera l'idée qu'il a dans sa tête et sera un point de départ pour la recherche dans la grande base de vidéos.

- Perspective 2

Exprimer un besoin de recherche par du texte est toujours beaucoup plus simple principalement pour les utilisateurs non informaticiens. Nous allons essayer dans des travaux futurs de permettre à un utilisateur de rechercher dans une grande base de vidéos en exprimant sa requête par du texte. Nous allons essayer d'annoter d'une manière automatique l'ensemble des images clés extraites en utilisant l'apprentissage profond afin de répondre à ce besoin.

RÉFÉRENCES

[Agarwal, 2008] Agarwal. G., Kempe. D. (2008). Modularity-Maximizing Graph Communities Via Mathematical Programming. The European Physical Journal. B, 66 (3).

[Alcantarilla, 2013] Alcantarilla. P. F., Bergasa. L. M., and Davison. A. 1. (2013). Gauge-SURF descriptors. Image and Vision Computing, 31.1, 103-116.

[Asadi, 2012] Asadi. E. and Charkari. N. M. (2012). Video Summarization Using Fuzzy CMeans Clustering. In 20th Iranian Conference on Electrical Engineering, (ICEE2012). IEEE, 690 – 694.

[Awad, 2003] Shao, H., Svoboda, T., Van Gool, L. (2003). Zubud-zurich buildings database for image based recognition. Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, Tech. Rep, 260, 20.

[Awad, 2016] Awad, A. I., et Hassaballah, M., (2016) Image Feature Detectors and Descriptors.

[Baber, 2013] Baber. 1., Afzulpurkar. N., and Satoh. S. 1. (2013). A framework for video segmentation using global and local features. International Journal of Pattern Recognition and Artificial Intelligence, 27(05).

[Barbieri, 2014] Barbieri. T. T. S., and Rudinei. G. (2014). KS-SIFT: A Keyframe Extraction Method Based on Local Features. (English) ISM: IEEE International Symposium on Multimedia, 13-17.

[Bay, 2008] Bay. H, Ess. A., Tuytelaars. T., Van Gool. L. (2008). SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, 346—359.

[Bahroun, 2011] Bahroun. S,. (2011). Construction de thésaurus pour la

recherche interactive dans une base d'images satellitaires. Thèse de doctorat.

[Binford, 1993] Binford. T.O., and Levitt. T.S. (1993). Quasi-invariants: theory and exploitation. Proceedings of Darpa Image Understanding Workshop, 819-829.

[Boukadida, 2015] Boukadida, H. (2015). Création automatique de résumés vidéo par programmation par contraintes (Doctoral dissertation, Université Rennes 1).

[Cai, 2005] B. Cai, Lu, Z., et Dong-ru, Z. (2005). A study of video scenes clustering based on shot key frames. Wuhan University Journal of Natural Sciences, 10(6), 966-970.

[Chandrasekhar, 2009] Chandrasekhar, V., Takacs, G., Chen, D., Tsai, S., Grzeszczuk, R., et Girod, B. (2009, June). CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 2504-2511).

[Chen, 1996] Chen. Y., and Hung. Y. (1996). Feature-based displacement field estimation for visual tracking by using coarse-to-fine block matching. International Conference on Articial Intelligence, 129-136.

[Chergui, 2012] Chergui. A., Bekkhoucha. A., and Sabbar. W. (2012). Video scene segmentation using the shot transition detection by local characterization of the points of interest. 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 404–411.

[Chun, 2001] Chi-Chun. Lo., Shuenn-Jyi. Wang. (2001). Video segmentation using a histogram-based fuzzy c-means clustering algorithm. Fuzzy Systems, 2001. The 10th lEEE international Conference on, vol.2, no., pp.920,923, vol.3, 2-5.

[Ciocca, 2006] Ciocca. G., and Schettini. R. (2006). An innovative algorithm

for key frame extraction in video summarization. In Journal of Real-Time Image Processing (in Print).

[Dang, 2015] Dang. C., Moghadam. A., Radha. H. (2015). RPCA-KFE: Key Frame Extraction for Consumer Video based Robust Principal Component Analysis. IEEE Transactions on Image Processing, 24(11).

[Daudin, 2008] Daudin. J.J., Picard. F., and Robin. S. (2008). A Mixture Model For Random Graphs. Statistics And Computing, 18:173–183.

[Deepak, 2013] Deepak. C.R., Babu. R.U., Kumar. K.8., Krishnan. C.M.R. (2013). Shot boundary detection using color correlogram and Gauge-SURF descriptors. Computing, Communications and Networking Technologies (ICCCNT), Fourth International Conference on, vol., no., pp. 1,5, 4-6.

[Dixit, 2013] Dixit. N. et Pro. Sandeep. Tiwari. (2016). International Journal of Advanced Research in Computer Science and Software Engineering, 6(6), pp. 84-89

[Dong, 2013] Dong. P. T. (2013). A Review on Image Feature Extraction and Representation Techniques. International Journal of Multimedia and Ubiquitous Engineering, Vol. 8, No. 4.

[Ejaz, 2012] Ejaz, N., Tariq T. B., and Baik, S. W., (2012). Adaptive key frame extraction for video summarization using an aggregation mechanism. Journal of Vision Communication and Image Representation, 23, 1031-1040.

[Fortunato, 2010] Fortunato. S. (2010). Community detection in graphs. Physics Reports, 486:75–174.

[Furini, 2010] Furini, M., Geraci, F., Montangero, M., Pellegrini, M. (2010). STIMO: STIll and MOving video storyboard for the web scenario. Multimedia Tools and Applications, 46(1), 47.

[Gil, 2010] Gil. A., Mozos. O.M., Ballesta. M. et al. (2010). A Comparative Evaluation of Interest Point Detectors and Local Descriptors for Visual SLAM", Machine Vision and Applications.

[Girgensohn, 2000] Girgensohn. A., Boreczky. J. (2000). Time-Constrained Keyframe Selection Technique. Multimedia Tools and Application, 11, 347-358.

[Gong, 200] Gong. Y., and Liu. X. Generating optimal video summaries. (2000). In IEEE International Conference on Multimedia and Expo (ICME'00), volume 3, 1559–1562, New York, USA.

[Gong, 2001] Gong. Y. and Liu. X. (2001). Video summarization with minimal visual content redundancies. In IEEE International Conference on Image Processing (ICIP'01), volume 3, 362–365, Thessalonique, Grèce.

[Grand-brochier, 2010] Grand-brochier. M., Tilmant. C. et Dhome. M. (2010). Combinaison du détecteur de points fast-hessien avec un descripteur local basé C-HOG. MajecSTIC.

[Grand-brochier, 2011] Grand-brochier. Manuel. (2011). Descripteurs 2D et 2D+t de points d'intérêt pour des appariements robustes. Thèse de doctorat, Université Blaise Pascal - Clermont II.

[Gunsel, 1997] Gunsel. B., Fu. Y., and Tekalp. A. M. (1997). Hierarchical temporal video segmentation and content characterization. In Proc SPIE Multimedia Storage and Archiving Systems II, volume 3229, 46–56,.

[Gygli, 2014] Gygli, M., Grabner, H., Riemenschneider, H., et Van Gool, L. (2014, September). Creating summaries from user videos. In European conference on computer vision (pp. 505-520). Springer, Cham.

[Hafiane, 2006] Hafiane. A., Zavidovique. B. (2006). Local relational string for textures classification. In IEEE International Conference on Image Processing, 21572160.

[Hannane, 2016] Hannane. R., Elboushaki. A., Afdel. K., Naghabhushan. P. and Jave. M. An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram. (2016). International Journal of Multimedia Information Retrieval, Volume 5, Issue 2,

89–104.

[Harris, 1988] Harris, C., et Stephens, M. (1988). A combined corner and edge detector. In Alvey vision conference (Vol. 15, No. 50, pp. 10-5244).

[Harwood, 1993] Harwood. D., Ojala. T., Pietikainen. M., Kelman. S., Davis. S. (1993). Texture classification by center- symmetric auto-correlation, using Kullback discrimination of distributions. Technical report, Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, Maryland. CAR-TR-678.

[Huang, 2017] Huang, X., Xu, Y., et Yang, L. (2017). Local visual similarity descriptor for describing local region. In Ninth International Conference on Machine Vision (ICMV 2016) (Vol. 10341, p. 103410S). International Society for Optics and Photonics.

[Hummel, 1983] Hummel and Zucker. S. (1983). On the fundations of relaxation labeling processes. IEEE Pattern Analysis and Machine Intelligence, 5(3), 267-287.

[Juan, 2010] Juan, L., et Gwun, O. (2010). A comparison of sift, pca-sift and surf. International Journal of Image Processing (IJIP), 3(4), 143-152.

[Karami, 2017] Karami, E., Prasad, S., et Shehata, M. (2017). Image matching using SIFT, SURF, BRIEF and ORB: Performance comparison for distorted images. arXiv preprint arXiv:1710.02726.

[Kingsbury, 2010] Kingsbury. NG., and Ng. ES. (2010). Matching of Interest Point Groups with Pairwise Spatial Constraints, 2693-2696.

[Krig, 2016] Krig, S. (2016). Interest point detector and feature descriptor survey. In Computer Vision Metrics (pp. 187-246). Springer, Cham.

[Kumar , 2011] Kumar. M., and Loui. A. C. (2011). Key Frame Extraction from Consumer Videos Using Sparse Representation. Proceedings of the 18th IEEE International Conference on Image Processing (ICIP2011), 2437-2440.

[Lai, 2012] Lai. J. L., and Yi. Y. (2012). Key frame extraction based on visual attention model. Journal of Vision Communication and Image Representation, 23, 114-125.

[Leordeanu, 2007] Leordeanu. M., and Hebert. M. (2007). A spectral technique for correspondence problems using pairwise constraints. ICCV, 1482–1489.

[Li, 2001] Li. Y., Zhang. T., and Tretter. D. (2001). An overview of video abstraction techniques. Technical report. HP Laboratory Technical Report.

[Li, 2009] Li, Y. N., et Lu, Z. M. (2009, December). Video abstraction via attention model and on-line clustering. In Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on (pp. 627-630). IEEE.

[Li, 2011] Li, Y., Merialdo, B., Rouvier, M., et Linares, G. (2011, November). Static and dynamic video summaries. In Proceedings of the 19th ACM international conference on Multimedia (pp. 1573-1576). ACM.

[Li, 2012] Li. J. (2012).Video Shot Segmentation and Key Frame Extraction Based on SIFT Feature. In 2012 International Conference on Image Analysis and Signal Processing (IASP). IEEE, 9-11, 1–8.

[Li, 2015] Li, J., et Wang, G. (2015, May). An improved SIFT matching algorithm based on geometric similarity. In Electronics Information and Emergency Communication (ICEIEC), 2015 5th International Conference on (pp. 16-19). IEEE.

[Liu, 2010] Liu, D., Hua, G., et Chen, T. (2010). A hierarchical visual model for video object summarization. IEEE transactions on pattern analysis and machine intelligence, 32(12), 2178-2190.

[Longfei, 2008] Longfei, Z., Yuanda, C., Gangyi, D., et Yong, W. (2008, December). A computable visual attention model for video skimming. In Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on (pp.

667-672). IEEE.

[Lowe, 2013] Lowe. D. G. (2004). Distinctive image features from scale-invariant keypoints. In. J.Comput. Vision, 91–110.

[Lu, 2003] Lu. Ye., Zhang. Hongjiang., Liu. W., and Hu. C. (2003). Joint semantics and feature based image retrieval using relevance feedback. IEEE Transactions on Multimedia, 5(3), 339–347.

[Lu, 2010] Lu. Bei., and Li. Qiuhong. (2010). Video key-frame-extraction based on block localfeatures and mean shift clustering. Wireless Communications Networking and Mobile Computing, 1-4.

[massaoudi, 2017] Massaoudi, M., Bahroun, S., Zagrouba, E., (2017). Video Summarization Based On Local Features. In : International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision.

[Mikolajczyk, 2005] Mikolajczyk, K., et Schmid, C. (2005). A performance evaluation of local descriptors. IEEE transactions on pattern analysis and machine intelligence, 27(10), 1615-1630.

[Morel, 2009] Morel, J. M., et Yu, G. (2009). ASIFT: A new framework for fully affine invariant image comparison. SIAM journal on imaging sciences, 2(2), 438-469.

[Mundur, 2006] Mundur, P., Rao, Y., Yesha, Y. (2006). Keyframe-based video summarization using Delaunay clustering. International Journal on Digital Libraries, 6(2), 219-232.

[Nagasaka, 1991] Nagasaka. A., and Tanaka. Y. (1991). Automatic video indexing and full-video search for object appearances. In Visual Database Systems, 113–127, Budapest, Hungary.

[Newman, 2004] Newman. M.E. and Girvan. M. (2004). Finding And Evaluating Community Structure In Networks. Physical Review, E, 69:026113.

[Ojala, 2000] Ojala. T., Pietikainen. M., Maenpaa. T. (2000). Multiresolution gray-scale and rotation invariant texture classification with local binary

patterns.

[Pan, 2013] Pan. B., Wang. Z. (2013) .Recent progress in digital image correlation. In: Application of Imaging Techniques to Mechanics of Materials and Structures, Volume 4, Conference Proceedings of the Society for Experimental Mechanics Series. Springer, New York, 317–326.

[Panchal, 2013] Panchal, P. M., Panchal, S. R., et Shah, S. K. (2013). A comparison of SIFT and SURF. International Journal of Innovative Research in Computer and Communication Engineering, 1(2), 323-327.

[Pang, 2012] Pang, Y., Li, W., Yuan, Y., et Pan, J. (2012). Fully affine invariant SURF for image matching. Neurocomputing, 85, 6-10.

[Pentland, 1994] Pentland. A., Picard. R., Davenport. G. and Haase. K. (1994). Video and Image Semantics: Advanced Tools for Telecommunications. IEEE MultiMedia, 1(2), 73-75.

[Rossi, 2012] Rossi. F., Villa-Vialaneix. N. (2012). Représentation D'un Grand Réseau A Partir D'une Classification Hiérarchique De Ses Sommets. Large Graph Visualization From A Hierarchical Node Clustering.

[Rosten, 2006] Rosten, E., et Drummond, T. (2006, May). Machine learning for high-speed corner detection. In European conference on computer vision (pp. 430-443). Springer, Berlin, Heidelberg.

[Rui, 1998] Rui. Y., Huang T. S. and Mehrotra. (1998). Relevance feedback techniques in interactive content-based image retrieval. IEEE Transactions On Circuits And Video Technology.

[Safavian, 1997] Safavian, S. R., Rabiee, H. R., et Fardanesh, M. (1997). Projection pursuit image compression with variable block size segmentation. IEEE Signal Processing Letters, 4(5), 117-120.

[Sandra, 2014] Sandra. E. F. de Avila., Ana. P. B. Lopes., Antonio. da Luz. Jr., Arnaldo. de A. Araújo. (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method Pattern

Recognition Letters. Volume 32, Issue 1, 56–68.

[Sargent, 2015] Sargent. G.,Perez-Daniel. KR., Stoian. A., Benois-Pineau. J., and Maabout. S. (2015). A scalable summary generation method based on cross-modal consensus clustering and OLAP cube modeling. Multimedia Tools and Applications, 1-22.

[Schaeffer, 2007] Schaeffer. S.E. (2007). Graph Clustering. Computer Science Review, 1(1):27–64.

[SenGupta, 2015] SenGupta. A. Video Shot Boundary Detection: A Review. (2015). IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).

[Shekar, 2011] Shekar. B. H., SharmilaKumari. M., and RaghuramHolla. (2011). Shot boundary detection algorithm based on color texture moments. Computer Networks and Information Technologies. Springer Berlin Heidelberg, 591-594.

[Shroff, 2010] Shroff, N., Turaga, P., et Chellappa, R. (2010). Video précis: Highlighting diverse aspects of videos. IEEE Transactions on Multimedia, 12(8), 853-868.

[Sidibe, 2007] Sidibe. D., Montesinos. P., and Janaqi. S. (2007). Mise en correspondance robuste d'invariants locaux par relaxation. ORASIS.

[Smith, 1997] Smith, S. M., et Brady, J. M. (1997). SUSAN—a new approach to low level image processing. International journal of computer vision, 23(1), 45-78.

[Sun, 2000] Sun. X., and Kankanhalli. M. S. (2000). Video Summarization Using RSequences. Real-Time Imaging, vol. 6, 449–459.

[Swain, 1991] Swain, M.J., Ballard, D.H., 1991. Color indexing. Internat. J. Comput. Vision 7 (1), 11– 32.

Pietikäinen, M., Hadid, A., Zhao, G., et Ahonen, T. (2011). Local binary patterns for still images. In Computer vision using local binary patterns

(pp. 13-47). Springer London.

[Takamura, 2011] Takamura, H., Yokono, H., et Okumura, M. (2011, April). Summarizing a document stream. In European conference on information retrieval (pp. 177-188). Springer, Berlin, Heidelberg.

[Tang, 2012] Tang, A., et Boring, S. (2012, May). EpicPlay: Crowd-sourcing sports video highlights. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1569-1572). ACM.

[Tapu, 2011] Tapu. R., and Zaharia. T. (2011). A complete framework for temporal video segmentation. Consumer Electronics - Berlin (ICCE-Berlin), IEEE International Conference, 156–1603.

[Thounaojam, 2014] Thounaojam. D. M., Trivedi. A., Singh. K. M., and Roy. S. (2014). A Survey on Video Segmentation. In intelligent Computing, Networking, and informatics, 903-912. Springer India.

[Tola, 2009] Tola, E., Lepetit, V., et Fua, P. (2010). Daisy: An efficient dense descriptor applied to wide-baseline stereo. IEEE transactions on pattern analysis and machine intelligence, 32(5), 815-830.

[Tono, 1993] Tonomura. Y., Akutsu. A., Otsugi. K., and Sadakata. T. (1993). VideoMAP and VideoSpaceIcon : Tools for automatizing video content. Proc. Acm Interchi Conference, 131-141.

[Tonomura, 1993] Tonomura. Y., Akustsu. A., Otsuji. K., and Sadakata. T. (1993). Videomap and video spaceicon: Tools for anatomizing video content. In Proceedings of the SIGCHI conference on Human factors in computing systems, 131–136.

[Torresani, 2008] Torresani. L., Kolmogorov. V., and Rother. C. (2008). Feature correspondence via graph matching: Models and global optimization. Proc. European Conference on Computer Vision, 596–609.

[Tsai, 2013] Tsai, C. M., Kang, L. W., Lin, C. W., et Lin, W. (2013). Scene-based movie summarization via role-community networks. IEEE Transac-

tions on Circuits and Systems for Video Technology, 23(11), 1927-1940.

[Tu, 1999] Tu. P., Saxena. T., and Hartley. R. (1999). Recognising objects using colour-annotated adjacency graphs. Lecture Notes in Computer Science; Shape, Contour and Grouping in Computer Vision, 246–263.

[Ueda, 1991] Ueda. H., Miyatake. T., and Yoshizawa. S. (1991). Impact: An interactive natural-motionpicture dedicated multimedia authoring system. In Proceedings of the SIGCHI conference on Human factors in computing systems, 343–350.

[Vinicius, 2017] Vinicius. M., Helio Pedrini. C. (2017). VISCOM: A robust video summarization approach using color co-occurrence matrices. Multimedia Tools and Applications, pp 1–19.

[Yuheng, 2017] Yuheng, S., et Hao, Y. (2017). Image Segmentation Algorithms Overview. arXiv preprint arXiv:1707.02051.

[Zabih, 1999] Zabih, R., Miller, J., Mai, K. (1999). A feature-based algorithm for detecting and classifying production effects. Multimedia systems, 7(2), 119-128.

[Zanghi, 2008] Zanghi. H., Ambroise. C., and Miele. V. (2008). Fast Online Graph Clustering Via Erdös-Rényi Mixture. Pattern Recognition, 41:3592–3599.

[Zhang, 1995] Zhang. H. J., Low. C. Y., Smoliar. S. W., and Zhong. D. (1995). Video parsing, retrieval and browsing: An integrated and content-based solution. In Proceedings of the 3rd ACM 194 BIBLIOGRAPHIE International Conference on Multimedia, 15–24, San Francisco, California, USA.

[Zhao, 2008] Zhao. Huan., Li. Xiuhuan., Yu. Lilei. (2008). Shot Boundary Detection Based on Mutual Information and Canny Edge Detector. Computer Science and Software Engineering, 2008, International Conference on , vol.2, no., pp.l124,1128, 12-14.

[Zhonghua, 2004] Zhonghua. Sun., Fu. Ping. (2004). Combination of Color and Object Outline Based Method in Video Segmentation. Proc. SPIE Storage and Retrieval Methods and Applications for Multimedia.