République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Ferhat Abbas de Sétif-1-

Thèse

Présentée à la Faculté des Sciences
Département d'informatique
En vue de l'obtention du diplôme de

Doctorat en sciences

Option : Informatique
Par

Abderraouf BOUZIANE

Thème

Organisation et indexation des données multimédias de grande dimension

Soutenue le 03 juillet 2013 devant le jury composé de :

Président	Dr. Abdallah Boukerram	Professeur à l'Université de Sétif-1-
Rapporteur	Dr. Messaoud Mostefai	MCA à l'Université de Bordj Bou Arréridj
Co-rapporteur	Dr. Youssef Chahir	HDR à l'Université de Caen, France
Examinateur	Dr. Mohamed-khireddine Kholladi	Professeur à l'Université d'El-Oued
Examinateur	Dr. Abdelouahab Moussaoui	Professeur à l'Université de Sétif-1-
Examinateur	Dr. Abdelhak Boubatra	MCA à l'Université de Bordj Bou Arréridj

Remerciements

Ce travail a été réalisé sous la direction de Monsieur Messaoud Mostefai, MCA à l'Université de Bordj Bou Arréridj. Qu'il trouve ici, l'expression de mes remerciements, mon profond respect et sincères considérations pour ses orientations, encouragements et son soutien pour faire aboutir ce travail de recherche à son terme.

J'ai été co-encadré par Monsieur Youssef Chahir, HDR à l'Université de Caen. Il m'a accueilli à Caen pendant 18 mois, la durée de ma Bourse PROFAS. A Caen, Youssef et sa famille étaient ma deuxième famille. Je ne saurais trouver les mots adéquats pour le remercier. Qu'il trouve ici, l'expression de mes remerciements, mon profond respect et sincères considérations pour ses orientations, encouragements et son soutien pour faire aboutir ce travail de recherche à son terme.

Je remercie Messieurs :

- Abdallah Boukerram, Professeur à l'Université de Sétif,
- Mohamed-khireddine Kholladi, Professeur à l'Université d'El-Oued,
- Abdelouahab Moussaoui, Professeur à l'Université de Sétif et
- Abdelhak Boubatra, MCA à l'Université de Bordj Bou Arréridj

pour avoir accepté de juger cette thèse de Doctorat et d'y apporter leur contribution scientifique.

Je remercie le Directeur du GREYC-Caen et le Responsable de l'équipe Image pour m'avoir accueilli au sein de leur Laboratoire.

Je remercie le Ministère de l'enseignement supérieur Algérien et le Ministère de l'enseignement supérieur Français pour m'avoir attribué une Bourse PROFAS afin de mener à terme ma thèse.

Je tiens à remercier particulièrement mon Ami Dr. Farid Nouioua, MCF à l'Université de Marseille, pour sa disponibilité, son encouragement et ses relectures attentives de mes articles et du présent manuscrit.

خلاصة

في الآونة الأخيرة، أ قترح النشر الهندسي باستخدام السير العشوائي على المخططات لإيجاد وصف هندسي معبر للبيانات وذلک من أجل تنظيمها.

الفكرة الأساسية هي في إعتبار مخطط لتمثيل جميع الأشياء المهمة (صورة، وثي قة، إلخ) في قاعدة البيانات وذلک من خلال أوصافها. تمثل قمم المخطط الأشياء المهمة ووزن الحواف ي قيس التشابه بين الأشياء. وعليه، يمكن اعتبار البحث عن شيئ موضع اهتمام كسير على المخطط. هذه الم قاربة تأخذ فى الاعتبار جميع الخصائص الاحصائية المتصلة بالمخطط لتوصيف أفضل للمعلومات البصرية. وعليه، يمكن للمستخدم تصفح قاعدة البيانات وعرض الصور المشابهة للصورة المرجعية.

في هذه الأطروحة ن قترح إطار موحد لتنظيم البيانات متعددة الوسائط ذات الأبعاد الكبيرة باستخدام هذه الم قاربة الجديدة. هذا الإطار يتكون من جزئين رئيسيين :

الجزء 1: التجميع الطيفي باستخدام السير العشوائي على المخطط.
الجزء 2: الترتيب المت قطع على المخطط.

ل قد جربنا وتح ق قنا من صحة هذا الإطار على البيانات الح قي قية، وذلک بعد مراحل من التصميم و التحليل النظريين.

الكلمات المفتاحية :

البيانات ذات الأبعاد الكبيرة، التجميع الطيفي، التعلم باستخدام المخططات، الترتيب المت قطع، السير العشوائي على المخططات، خرائط النشر، تحليل السلوک البشري باستخدام الفيديو، التجزئة الهرمية لدفق الفيديو، أعزام ZERNIKE الزمانية المكانية، الن قاط SIFT.

Résumé

Récemment, la diffusion géométrique par marches aléatoires sur graphe est proposée pour trouver une description géométrique significative des données afin de les organiser.

L'idée de base est de considérer un graphe modélisant tous les objets d'intérêt (image, document, etc.) d'une base à travers leurs descripteurs. Les sommets représentent les objets d'intérêt et la pondération des arêtes mesure la similarité entre objets. Ainsi, la recherche d'un objet d'intérêt peut être vue comme une marche dans ce graphe. Cette approche permet de tenir compte de toutes les propriétés statistiques liées au graphe local pour caractériser au mieux l'information visuelle. Ainsi, l'utilisateur peut naviguer dans la base et visualiser les images voisines à une image de référence.

Dans cette thèse de Doctorat nous proposons un framework unifié pour l'organisation des données multimédias de grande dimension en utilisant cette nouvelle approche. Ce framework se compose de deux grandes parties :

- Partie 1 : regroupement spectral par marches aléatoires sur graphe.
- Partie 2 : régularisation discrète sur graphe.

Après une phase de conception et d'analyse théorique, nous avons expérimenté et validé ce framework sur des données réelles.

Mots-Clés :

Données de grande dimension, Regroupement spectral, Apprentissage en utilisant les graphes, Régularisation sur graphes, Marches aléatoires sur graphes, Diffusion maps, Analyse vidéo du comportement humain, Segmentation hiérarchique des flux vidéo, Moments de Zernike spatio-temporels, Points SIFT.

Abstract

Recently, the geometric diffusion by using random walks on graph is proposed to find a significant geometric description of data and to organize them.

The basic idea is to consider a graph modeling all objects of interest (image, document, etc.) of a database through their descriptors. Vertices represent the objects of interest and the weight of the edges measure the similarity between objects. Thus, the search for an object of interest can be seen as a walk in this graph. This approach takes into account all statistical properties related to the graph to best characterize the visual information. Thus, the user can browse the database and view the images nearby to a reference image.

In this Thesis we propose a unified framework for the organization of large multimedia data using this new approach. This framework consists of two main parts :

- Part 1 : spectral clustering by random walks on graph.
- Part 2 : discrete regularization on graph.

After a phase of design and theoretical analysis, we have experimented and validated this framework on real data.

Key-words :

High-dimensional data, Spectral clustering, Graph-based learning, Regularization on graphs, Random walks on graphs, Diffusion maps, Video analysis of human behaviour, Hierarchical video stream segmentation, Spatio-temporal Zernike moments, SIFT features.

Table des matières

Table des figures

Introduction générale

Contexte de recherche

Dans un grand nombre d'applications en vision par ordinateur, recherche des images par le contenu, structuration des documents multimédias, etc., nous sommes souvent confrontés à des données massives de grandes dimensions (image, texte, vidéo, son, etc.), où chaque élément est décrit par un nombre d'attributs considérable et donc distribués dans un espace de grande dimension. Cette grande proportion représente généralement une entrave pour le traitement, l'organisation, la recherche, l'analyse ou la visualisation de ces données.

De plus, une des questions les plus pertinentes est la validité (fiabilité) des résultats de regroupement (clustering) et de classification des données de grande dimension. En effet, étant donné que les algorithmes de regroupement sont étroitement liés aux mesures de similarité et par conséquent, à la recherche des plus proches voisins d'un objet donné, cette recherche peut s'avérer inefficace car il est très difficile de faire la différence entre le voisin le plus proche et les autres voisins dans les espaces de grande dimension. Effectivement, il a été montré dans [Beyer, 1999], qu'au fur et à mesure que le nombre d'attributs nécessaires à la description d'un objet (la dimension de l'espace de données) augmente, la distance d'un objet à son plus proche voisin et sa distance à son voisin le plus éloigné ont tendance à converger.

Cependant, il est généralement observé que la "dimension intrinsèque" des données est souvent beaucoup plus faible que la dimension de l'espace où elles sont observées. Ainsi, la réduction de leur dimension, non seulement, permet d'avoir plus de certitude sur le résultat de leur classification mais permet, également, d'améliorer leur lisibilité.

Motivations

Il y a eu un intérêt considérable pour les techniques visant à réduire la dimension des données en cherchant des représentations appropriées, des données "complexes", qui permettent d'identifier les dimensions qui sont porteuses de l'information pertinente. L'objectif est d'identifier des structures de petite dimension cachées dans des observations de grande dimension tout en gardant le maximum d'informations sur la structure originale des données.

Plusieurs approches sont proposées pour la réduction de la dimensionnalité. Parmi celles-ci, **les méthodes spectrales** proposent d'utiliser les vecteurs et valeurs propres de diverses manières pour déduire la structure ainsi que les propriétés principales de l'espace réduit et cela en utilisant le spectre de l'espace d'origine (espace d'observation, appelé aussi espace de mesure).

Pour récupérer la topologie des données de l'espace observé, ces méthodes suivent, globalement, la même démarche.

- Chercher les plus proches voisins aux données en entrées.
- Construire un graphe de voisinage pondéré.
- Dériver une matrice (un noyau) à partir de ce graphe.
- Déduire une base orthonormée pour l'espace réduit en utilisant le spectre (vecteurs et valeurs propres) de cette matrice.

Récemment, la diffusion géométrique par marches aléatoires sur graphe est proposée pour trouver une description géométrique significative des données permettant de découvrir et d'organiser les données. En conséquence, cela permettra une réduction non linéaire de données nécessaire pour leur traitement, leur classification, voir leur visualisation.

L'idée de base est de considérer un graphe modélisant tous les objets d'intérêt (image, document, etc.) d'une base à travers leurs descripteurs. Les sommets représentent les objets d'intérêt et la pondération des arêtes mesure la similarité (ou dissimilarité) entre objets. Ainsi, la recherche d'un objet d'intérêt peut être vue comme une marche dans ce graphe. Cette approche permet de tenir compte de toutes les propriétés statistiques liées au graphe local pour caractériser au mieux l'information visuelle. Ainsi, l'utilisateur peut naviguer dans la base et visualiser les images voisines à une image de référence.

L'objectif de cette thèse de Doctorat est de proposer un framework pour l'organisation des données multimédias de grande dimension en utilisant cette nouvelle approche. Puis, de l'expérimenter et de le valider sur des données réelles et cela après une phase de conception et d'analyse théorique.

Contributions

Les contributions de cette thèse peuvent être situées sur quatre plans :

1. **Utilisation des moments de zernike 3D pour décrire la silhouette et la dynamique d'un objet d'intérêt dans un clip vidéo :**

   Dans cette thèse, nous proposons une nouvelle méthode de classification des actions humaines basée sur une extension des moments de Zernike au domaine spatio-temporelle et ce afin de caractériser les actions humaines dans les séquences vidéo. En effet, les moments sont largement utilisés dans les statistiques et dans le domaine de la reconnaissance des formes pour caractériser une fonction et d'en saisir les caractéristiques importantes, en particulier pour la description des formes. L'utilisation des moments de Zernike 3D, calculés pour des volumes d'espace-temps, a pour but de saisir des informations, à la fois structurelles et temporelles, d'une séquence variant dans le temps. Ces moments sont par construction non redondants et donc optimaux pour la compacité et pour la reconstruction.

2. **Définition d'un cadre général d'organisation et de diffusion des informations sur graphe :**

   Comment organiser et classer les données multimédias de grande dimension est un sujet de recherche important. Une solution à ce problème, non seulement, facilite les applications telle que la classification vidéo mais permet, également, d'améliorer, par exemple, des systèmes automatiques de vidéo surveillance et de communication humain-machine/robot.
   En plus de son importance pour de nombreuses applications pratiques, la catégorisation, en particulier non supervisée, des données multimédias est importante dans le contexte de l'apprentissage automatique. Elle permet de décrire la manière dont les approches de traitement des documents multimédias pourrait permettre une compréhension de haut niveau des données.
   Pour apporter une solution à ce sujet, nous proposons dans cette thèse un framework unifié pour la réduction et la catégorisation des données multimédias de grande dimension. Nous décrivons, entre autres, comment ce framework peut permettre un apprentissage faiblement supervisé à partir d'un sous ensemble de données.

   Le framework proposé se compose de deux grandes parties :

(a) **Partie 1 :** suivant la nature des données, des descripteurs adéquats sont extrait. Ensuite, nous construisons un graphe de similarité visuelle en calculant la similarité par paire entre ces descripteurs. L'idée de base est de considérer toutes les données multimédias comme un graphe pondéré dont les sommets (des données) représentent ces descripteurs et les arêtes représentent la similarités entre les sommets connectés. A ce stade, nous utilisons le spectre de la matrice associée aux marches aléatoires sur ce graphe pour définir de nouveaux descripteurs plus compactes des données multimédias de grandes dimension. Aussi, une nouvelle famille de distances est définie et sera utilisée pour décrire comment deux descripteurs sont similaires et par conséquent, catégoriser les données observées.

(b) **Partie 2 :** dans cette partie, nous décrivons une approche de régularisation discrète du graphe construit sur les nouveaux descripteurs dans l'espace réduit. Ce cadre de régularisation exploite la géométrie des données pour :

i. L'élimination des données aberrantes, et/ou

ii. La prédiction des données manquantes.

Avec ce cadre, nous pouvons considérer le problème général de l'apprentissage d'une fonctionnelle de dépendance inconnue entre un espace structuré, en entrée, et un autre espace structuré, en sortie, à partir d'exemples labellisés et non labellisés.

Ce framework est mis en œuvre à travers cinq modules :

- **Module de prétraitement :** à partir des objets de l'espace d'observation, des descripteurs adéquats sont calculés. Ces derniers offrent une discrimination aussi nette que possible des données en entrée et permettent le passage à un autre espace de représentation quantifiable. Plusieurs exemples ont été étudié, notamment le cas des images fixes et de la vidéo.
- **Module de calcul de similarité :** après avoir obtenu les descripteurs des objets d'intérêt, un graphe est utilisé pour modéliser le nouvel espace de représentation où les sommets représentent les objets d'intérêt et la pondération des arêtes mesure la similarité entre objets. Le calcul de la similarité passe d'abord par le calcul des distances où nous avons testé plusieurs métriques ensuite différentes variantes du graphe de similarité ont été étudiées.
- **Module des marches aléatoires sur graphe :** le graphe de similarité est utilisé pour générer la matrice associée à une chaîne de Markov. Une décomposition spectrale de cette dernière définit un espace de projection euclidien permet-

tant ainsi une réduction non linéaire des données en entrée. Les données issues du module de décomposition spectrale sont classées et une catégorisation de l'espace des données en entrée est proposée.

– **Module d'accélération des marches aléatoires sur graphe :** Avec ce module, seuls les premiers vecteurs et valeurs propres, de la matrice de transition, sont calculés. Comparé à la méthode SVD, cela permettra un gain de temps de calcul relativement important.
– **Module de régularisation discrète sur graphe :** afin d'améliorer la classification des données et réduire l'impact des valeurs aberrantes, ce module permet de débruiter l'espace réduit obtenu par le module de décomposition spectrale et permet aussi de diffuser les labels pour la prédiction des données manquantes.

3. **Accélération des marches aléatoires sur graphe :**

Les marches aléatoires sur graphe définissent une matrice de transition. Pour un ensemble de données de grande dimension, la solution de son équation caractéristique prend beaucoup de temps et ralentit, ainsi, le calcul des valeurs et vecteurs propres nécessaires à la définition des nouvelles coordonnées réduites.

Pour palier à ce problème, nous utilisons l'algorithme des puissances itérées et la méthode de déflation pour approximer les valeurs et vecteurs propres. Cela est effectué en cherchant les premiers vecteurs propres dominants et leurs valeurs propres correspondantes.

En effet, l'algorithme des puissances itérées est une méthode simple pour calculer le vecteur propre dominant car il accède à une matrice uniquement à travers sa multiplication par des vecteurs. Cette propriété est particulièrement importante dans le cas des matrice de très haut rang. La méthode de déflation, quant à elle, permet de retirer, à chaque itération, la valeur propre dominante et réarrange la matrice de telle sorte que la valeur propre dominante de cette nouvelle matrice n'est autre que la deuxième valeur propre de la matrice d'origine. Ce processus est répété pour calculer les valeurs propres une par une.

4. **Réalisation de trois applications pour l'évaluation et la validation du framework :**

Plusieurs applications ont été développées pour évaluer notre approche :

(a) Un cadre unifié pour la reconnaissance des comportements humains.

(b) Apprentissage faiblement supervisé à partir des points SIFT.

(c) Structuration des flux TV en utilisant un framework de diffusion.

Le schéma suivant illustre notre framework :

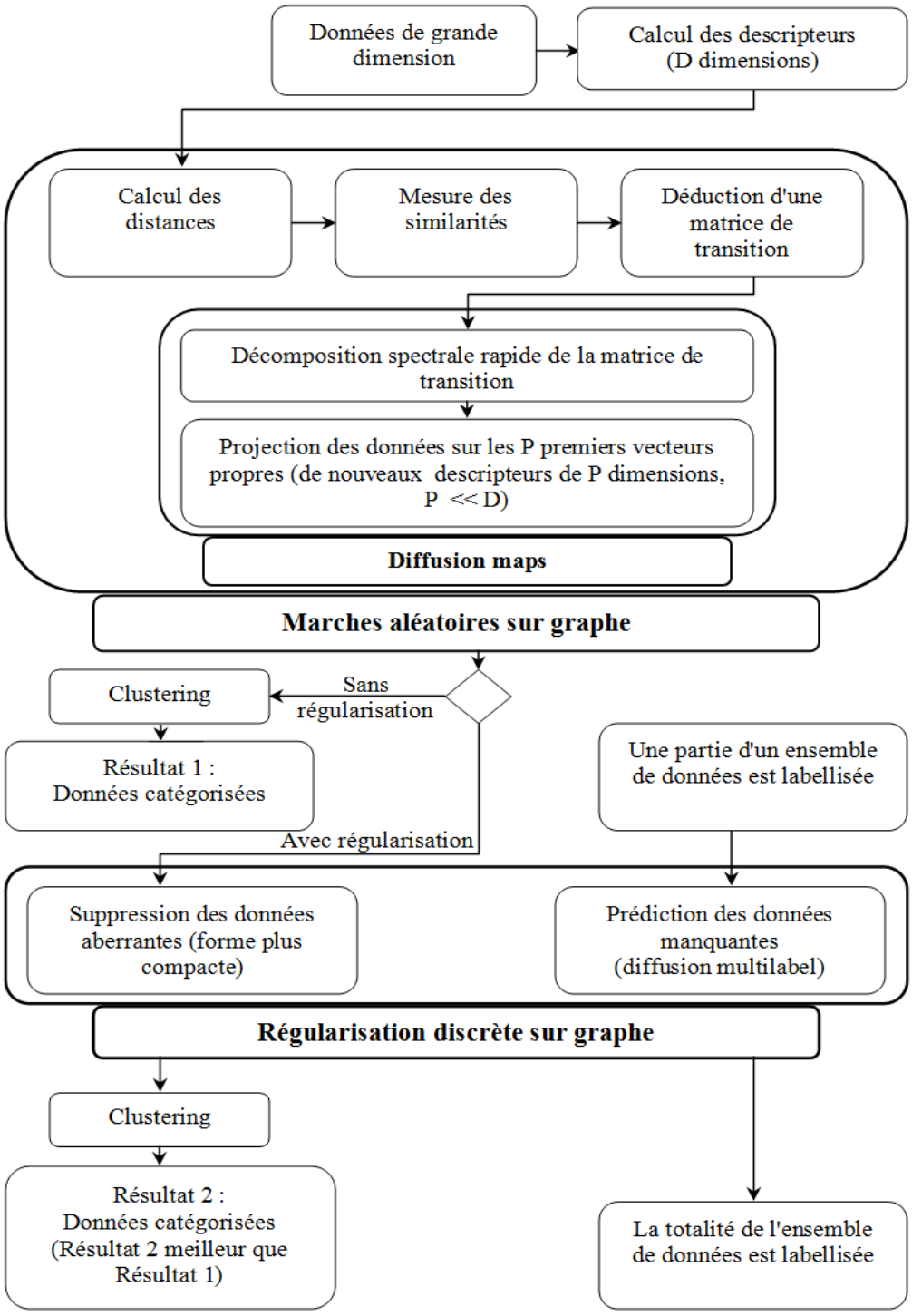

FIGURE 0.1 – Le framework proposé pour l'organisation des données de grande dimension

Plan du manuscrit

La présente thèse est structurée en deux parties. Chacune est subdivisée en trois chapitres. La première partie introduit la base théorique de notre framework. Elle regroupe les chapitres suivants :

- **Le chapitre 1 :** ce chapitre constitue l'état de l'art sur les méthodes spectrales pour le regroupement et la réduction des données qui peuvent être décomposées en deux grandes classes : les méthodes linéaires et les méthodes non linéaires. Nous décrirons, dans ce chapitre, les formes les plus simples de quelques algorithmes représentatifs.

- **Le chapitre 2 :** à ce niveau, nous détaillons l'approche de catégorisation de données que nous avons utilisée. Cette approche se base sur la diffusion géométrique par marches aléatoires sur graphe pour la définition d'une nouvelle famille de distances, appelée diffusion maps. Cette nouvelle approche fait partie des méthodes non linéaires pour la réduction de la dimensionnalité.
  A partir des données observées jusqu'à la définition des nouvelles coordonnées, dans l'espace réduit, nous montrerons d'abord :

  - Les différentes configurations possibles pour le graphe de voisinage.
  - Les différents descripteurs utilisés dans les applications développées dans la deuxième partie.
  - Les différentes distances utilisées.
  - Les différentes mesures de similarité (locale et non locale).

  Nous illustrerons les étapes nécessaires à la définition de la nouvelle famille de distances ainsi que les nouvelles coordonnées des données observées dans l'espace réduit. Nous détaillons, également, la méthode que nous avons utilisée pour l'accélération des marches aléatoire sur le graphe de données.

- **Le chapitre 3 :** ce chapitre est consacré à la définition du cadre général de régularisation discrète sur graphe. Nous introduirons, d'abord, quelques notions sur les graphes et définissons quelques opérateurs nécessaires au développement du cadre de régularisation. Ensuite, nous décrivons la méthode de régularisation.

Les concepts définis dans la première partie sont mis en œuvres à travers trois applications. Ces applications sont présentées dans les trois derniers chapitre qui composent la deuxième partie :

- **Le chapitre 4 :** dans ce chapitre, nous présentons un cadre unifié pour la reconnaissance des comportements humains. Le but de notre travail est de développer une approche visuelle pour la reconnaissance du comportement humain. Notre contribution est illustrée par deux applications :
  - Catégorisation des actions humaines.
  - Une approche visuelle pour l'identification des propriétés d'un objet manipulé par les mains.

  Ces deux applications utilisent une extension des moments de Zernike au domaine spatio-temporelle qui permet de caractériser les actions de l'être humain dans les clips vidéo.

- **Le chapitre 5 :** l'application développée dans ce chapitre traite l'apprentissage faiblement supervisé à partir des points SIFT. Nous proposons une approche unifiée de diffusion de labels dans un espace de grande dimension en utilisant un ensemble restreint de points SIFT. Ces points sont identifiés automatiquement et permettent la segmentation des images. Les résultats de la segmentation ont montré que notre approche permet de propager efficacement les connaissances initiales et d'obtenir, ainsi, de bons résultats. L'objectif de cette application est de montrer comment est-il possible d'apprendre à partir d'un ensemble restreint de données et de propager les connaissances acquises dans une base de données de grande dimension.

- **Le chapitre 6 :** ce chapitre décrit l'application dédiée à la structuration des flux TV en utilisant un framework de diffusion. Cette application permet, en particulier, de construire une hiérarchie des résumés du flux TV allant du plus détaillé au plus général. Par conséquent, le spectateur peut composer son propre programme en ignorant les résumés qui ne l'intéressent pas ou en naviguant en profondeur dans d'autres. Nous décrivons aussi notre algorithme de segmentation vidéo interactif multi-label qui propose au spectateur de retrouver, dans son programme, les résumés qui partagent le même contenu visuel.

Nous terminons cette thèse par une conclusion générale de notre travail suivie des perspectives ouvrant des horizons pour d'autres réalisations dans le même cadre.

Première partie

Regroupement spectral et régularisation discrète sur graphe

Chapitre 1

Méthodes de regroupement spectral pour la réduction de la dimension

Sommaire

1.1 Introduction

Les informations multimédias (texte, images, séquences vidéo,···) générées par les applications informatiques deviennent de plus en plus volumineuses et la recherche d'une information précise est devenue une tache difficile qui prend beaucoup de temps.

Afin de faciliter l'accès aux données, une approche commune consiste à structurer ces informations dans des bases de données où chacune de ces données est composée d'un ensemble d'objets. A son tour, un objet est décrit par l'intermédiaire d'un ensemble de caractéristiques (vecteur des caractéristiques) exprimées à l'aide des valeurs scalaires. Par exemple, dans une base d'images, ces caractéristiques peuvent être : les moments, les couleurs, les textures, les descripteurs de formes, etc. Cependant ce vecteur des caractéristiques comporte, souvent, plusieurs entrées et il est donc réparti dans un espace de grande dimension.

Ainsi, chercher une information particulière, revient à chercher la similarité entre son (ses) vecteur(s) caractérisant(s) et d'autres vecteurs contenus dans la base de données. Cette recherche de similarité (qui n'est autre que la recherche des plus proches voisins) devient de plus en plus difficile lorsque la dimension du vecteur des caractéristiques augmente. Par conséquent, le nombre d'objets de la base de données qui doivent être examinés augmente exponentiellement avec la dimension sous-jacente du vecteur des caractéristiques [Bellman, 1961].

Une des questions les plus pertinentes est la signification des mesures de similarité dans les espaces de grande dimension. Beyer et al. [Beyer, 1999] ont montré qu'au fur et à mesure que la dimension de l'espace de données augmente, la distance d'un objet à son plus proche voisin et sa distance à son voisin le plus éloigné, ont tendance à converger. Cela signifie que la recherche du plus proche voisin peut s'avérer inefficace car il est très difficile de faire la différence entre le voisin le plus proche et les autres voisins.

– Dimension intrinsèque des données

Dans de nombreuses situations, la "dimension intrinsèque" d'un ensemble de données est souvent beaucoup plus faible que la dimension de l'espace d'observation. Cela se produit principalement lorsque les valeurs de certaines caractéristiques sont corrélées (linéairement dépendantes). Dans d'autres cas, l'expert peut considérer que certaines entrées du vecteur des caractéristiques (fonctions) ne sont pas aussi importantes que les autres entrées dans la discrimination entre les objets et peuvent donc être ignorées ou affectées d'une faible pondération. Par exemple, si on considère une scène vidéo visualisant le déplacement d'un même individu dans la même zone alors,

quelque soit la taille de cette dernière, l'information à en extraire revient à considérer un certain nombre d'images auxquelles sont appliqués quelques déplacements horizontaux et verticaux. Ce qui revient à chercher un nombre minimal de variables nécessaires pour décrire la position de cet individu sans ambiguïtés .

Plus formellement, ayant un ensemble $X = (x_1, \ldots, x_n)$, de données de grande dimensions ($x_i \in \mathbb{R}^d$), le but est de trouver une représentation équivalente $Y = (y_1, \ldots, y_n)$ ($y_i \in \mathbb{R}^m$) où

1. $m << d$ et
2. les voisins proches dans X restent proches dans Y et ceux éloignés dans X le restent, aussi, dans Y.

Plusieurs approches de réduction de dimension existent. Parmi celles-ci, **les méthodes spectrales** proposent d'utiliser les vecteurs et valeurs propres de diverses manières pour déduire la structure ainsi que les propriétés principales de l'espace réduit à partir du spectre de l'espace d'origine (espace d'observation, appelé aussi espace de mesure). Dans ce qui suit, nous décrirons les formes les plus simples de quelques algorithmes représentatifs qui font partie de l'une des deux grandes classes : les méthodes linéaires et les méthodes non linéaires.

1.2 Les méthodes linéaires

La réduction des dimensions peut être très utile pour améliorer la lisibilité des données et cela en supprimant un grand nombre de dimensions. Pour atteindre cet objectif, la stratégie adoptée consiste à éliminer les variables moins pertinentes tout en minimisant l'erreur de reconstruction.

1.2.1 Analyse en composantes principales (ACP)

L'analyse en composantes principales (ACP) a été introduite, initialement, par Pearson [Pearson, 1901], et a subie plusieurs développements par la suite ([Hotelling, 1933, Karhunen, 1946, Loeve, 1948]). Dans le modèle de données ACP, on suppose qu'un vecteur de caractéristiques $\mathbf{y} = [y_1, y_2, \cdots, y_d]^T$ de d variables observées est le résultat d'une transformation linéaire W d'un vecteur $\mathbf{x} = [x_1, x_2, \cdots, x_p]^T$ de p variables latentes inconnues :

$$y = Wx \tag{1.1}$$

La matrice W et les p variables latentes (p correspond à la dimension de l'espace réduit) sont à déterminer sachant que les colonnes de la matrices W forment une base (sont linéairement indépendants) et que :

$$W^T W = I_p \tag{1.2}$$

Plusieurs critères permettent de définir le modèle ACP. Parmi lesquels, nous avons le critère de minimisation de l'erreur de reconstruction suivante :

$$\mathscr{E} = \left\| y_i - \sum_{j=1}^{p} (y_i \,.\, e_j)\, e_j \right\|^2 \tag{1.3}$$

où les vecteurs $(e_j)_{j=1}^{p}$ définissent une base partielle orthonormée de l'espace d'entrée (celui de d dimensions). Les vecteurs de la nouvelle base (de l'espace réduit) sont donnés par les p premiers vecteurs propres de la matrice de covariance ($d \times d$) suivante :

$$C = \frac{1}{n} \sum_i y_i y_i^T \tag{1.4}$$

Ainsi, les p premières valeurs propres significatives identifient la dimension du nouvel espace et les nouvelles coordonnées $\mathbf{x}$ sont données par :

$$x_{ij} = y_i \,.\, e_j \tag{1.5}$$

1.2.2 Métrique de positionnement multidimensionnel (MDS)

La méthode MDS cherche à définir la base d'un espace réduit à partir d'une population de N objets, où chacun d'eux est caractérisé (dans l'espace observé) par des distances (ou des similarités) envers les autres objets. La méthode MDS préserve les produits scalaires mutuels entre les objets au lieu d'en préserver les distances.

Comme dans le cas du ACP, MDS est basée sur un modèle linéaire simple où pour des valeurs observées $\mathbf{y} = [y_1, y_2, \cdots, y_d]^T$ de d dimensions, elle cherche à identifier les p axes pertinents qui définissent la base de l'espace réduit :

$$y = W x\,, \qquad (W^T W = I_p) \tag{1.6}$$

Pour un ensemble de N objets, la matrice ($N \times d$) :

$$\mathbf{Y} = [\cdots, y(i), \cdots, y(j), \cdots]^T \tag{1.7}$$

définit l'ensemble des distances mutuelles (similarités) entre ces objets. Le produit scalaire entre les vecteurs $y(i)$ et $y(j)$ est donné par :

$$s_{\mathbf{y}}(i,j) = s(y(i), y(j)) = \langle y(i) \,.\, y(j) \rangle \tag{1.8}$$

Donc, la matrice des produits scalaires, **S**, appelée matrice de Gram, est comme suit :

$$\mathbf{S} = [s_{\mathbf{y}}(i,j)]_{1 \leqslant i,j \leqslant N} = \mathbf{Y}^T\mathbf{Y} = (\mathbf{WX})^T(\mathbf{WX}) = \mathbf{X}^T\mathbf{W}^T\mathbf{WX} = \mathbf{X}^T\mathbf{X} \tag{1.9}$$

$$(W^T W = I_p)$$

Les nouvelles coordonnées $x_i \in \mathbb{R}^m$, obtenues par MDS, sont ceux qui minimisent :

$$\mathscr{E} = \sum_{ij} (y_i \,.\, y_j - x_i \,.\, x_j)^2 \tag{1.10}$$

et elles sont obtenues de la décomposition spectrale de la matrice **S**. Soit $(e_j)_{j=1}^p$ les p premiers vecteurs propres de la matrice **S**, et $(\lambda_i)_{i=1}^p$ leurs valeurs propres correspondantes. Les nouvelles cordonnées sont alors données par :

$$x_{ij} = \sqrt{\lambda_i}\, e_{ij} \tag{1.11}$$

Bien que basée sur une intuition géométrique un peu différente, MDS donne les mêmes résultats que l'ACP. Dans les deux techniques, un écart important entre la p-ième et la $(p+1)$ valeur propre indique que les coordonnées d'entrée (de grande dimension) sont très bien approximées par de nouvelles coordonnées de dimension largement inférieure.

1.3 Les méthodes non linéaires

Les méthodes linéaires supposent qu'une variable observée $\mathbf{y} = [y_1, y_2, \cdots, y_d]^T$ qui est difficile à interpréter vu sa grande dimension, n'est en fait qu'une expression linéaire de certaines autres valeurs latentes $\mathbf{x} = [x_1, x_2, \cdots, x_p]^T$ de dimension largement inférieure et qui sont, contrairement à **y**, facilement exploitables. Derrière cette hypothèse,

il y a la notion de réduction linéaire de dimension et ainsi que la notion de séparation de variables latentes (orthogonalité). Cependant, si les données sont réparties dans un espace non linéaire, ces méthodes ne permettent pas d'aboutir au résultat escompté.

Pour palier à ce problème, les méthodes non linéaires utilisent les graphes pour récupérer la topologie des données dans l'espace observé ainsi que leurs propriétés statistiques. Dans ce qui suit, nous présentons quelques algorithmes représentatifs.

1.3.1 Isomap

Cet algorithme [Tenenbaum, 1998, Tenenbaum, 2000] peut être vu comme une variante de la méthode MDS sur graphe, où la distance géodésique remplace la distance euclidienne entre les objets observés. Isomap s'exécute en trois étapes séquentielles qui sont :

1. Construire un graphe des k-plus proches voisins. Les sommets de ce graphe représentent les objets de l'espace observé. Le poids d'une arrête (non dirigée) représente la distance euclidienne entre les objets connectés.
2. Calculer, pour toutes les paires (u,v) des sommets du graphe, une nouvelle distance en utilisant un algorithme du plus court chemin. (l'algorithme de Djikstra par exemple).
3. Une fois cette matrice de distance calculée, utiliser l'algorithme MDS (les nouvelles distances en constitueront l'entrée) pour déduire les nouvelles coordonnées des objets observés dans l'espace réduit.

Isomap permet d'obtenir une nouvelle représentation de faible dimension dans laquelle les distances euclidiennes entre les objets dans l'espace réduit correspondent aux distances géodésiques entre les objets observés.

1.3.2 Locally Linear Embedding

L'algorithme "Locally linear embedding" (LLE) [Roweis, 2000] cherche à définir un espace de faible dimension à partir d'un espace observé de grande dimension et de topologie inconnue. Pour ce faire, l'idée de base est de découper l'espace en entrée en des sous-espaces suffisamment petits pour qu'ils soient considérés linéaires et exprimer, ainsi, les variables observées dans ces sous-espaces par des combinaisons linéaires.

Formellement, LLE agit en trois étapes :

1. Trouver les k-plus proches voisins pour chaque variable observée y_i. Le graphe à construire dans ce cas est orienté. Ces arcs indiquent les relations de voisinage

(k-pp) qui peuvent être symétriques ou pas. Soit $\mathcal{N}(i)$ le voisinage de y_i.

2. Affecter des poids w_{ij} aux arêtes de ce graphe. Les poids w_{ij} sont calculés en reconstruisant chaque variable y_i à partir de ses k-plus proches voisins y_j. Ils sont choisis pour minimiser l'erreur de reconstruction suivante :

$$\mathscr{E} = \left\| y_i - \sum_j (w_{ij}\ y_j) \right\|^2 \tag{1.12}$$

Cette minimisation est effectuée sous les conditions :

(a) $w_{ij} = 0$ si $x_j \notin \mathcal{N}(i)$

(b) $\Sigma_j w_{ij} = 0,\ \forall i$

La matrice des poids W permet ainsi de capturer les propriétés géométriques locales de l'espace observé en spécifiant la relation de chaque variable y_i avec ces k-plus proches voisins.

3. Déduire la projection x_i des objets observés y_i dans l'espace réduit en minimisant l'erreur de reconstruction suivante :

$$\mathscr{E} = \left\| x_i - \sum_j (w_{ij}\ x_j) \right\|^2 \tag{1.13}$$

Ils sont obtenus en calculant les $(p+1)$ derniers vecteurs et valeurs propres de la matrice $(I-W)^T(I-W)$ (le dernier vecteur propre étant constant et par conséquence, il est ignoré).

1.3.3 Laplacian Eigenmaps

La méthode Laplacian Eigenmaps a été introduite par Belkin et al. [Belkin, 2003]. Elle diffère de LLE dans la façon de construction de la matrice des poids W. Au lieu de reproduire des sous-espaces linéaires autour de chaque donnée, Laplacian Eigenmaps est basée sur la minimisation des distances entre les données (points) voisines (distance locales) afin de préserver la topologie de l'espace observé et éviter ainsi que tous les points soient projetés sur un seul point.

Aussi, Laplacian Eigenmaps est compsée de trois étapes, qui sont :

1. Construire un graphe de voisinage (k-pp par exemple).
2. Affecter à chaque arrête (y_i, y_j), le poids $w_{ij} = exp(-\|y_i - y_j\|^2/\sigma^2)$. σ étant un paramètre d'échelle.
3. Définir les coordonnées de l'espace réduit en utilisant les $(p+1)$ derniers vecteurs et valeurs propres de la matrice $\mathscr{L} = I - D^{-1/2}\ W\ D^{-1/2}$. D est la matrice diagonale où $D_{ii} = \Sigma_j w_{ij}$.

Chapitre 2

Marches aléatoires sur graphe pour structurer les données de grande dimension

Sommaire

2.1 Introduction

L'idée de base de l'approche, de structuration des données multimédias de grande dimension, que nous avons mise en œuvre consiste à considérer un graphe modélisant tous les objets d'intérêt (vidéos, images, documents, etc.) d'une base à travers leurs descripteurs. Les sommets représentent les objets d'intérêt et la pondération des arêtes mesure la similarité entre ces objets. Cette technique permet :

- d'associer aux données multimédias, une description conservant leur cohérence présente dans la séquence à traiter,
- de capturer les relations qui gouvernent ces objets, et moyennant le spectre de ce graphe,
- d'extraire un code contenant l'information pertinente en exploitant la redondance des descripteurs.

Les informations pertinentes, ainsi déduites, sont une forme plus compacte des données originales. De plus, les nouvelles coordonnées trouvées permettent le passage d'un espace de grande dimension et de topologie inconnue à un espace euclidien de dimension largement inférieure.

– Préliminaires

Soit $G = (V, E, w)$ un graphe valué non orienté où $V = \{v_1, v_2, \ldots, v_n\}$, est un ensemble fini de sommets représentant un ensemble fini de données et $E \subseteq V \times V$, un ensemble fini d'arêtes représentant la similarité entre les sommets connectés. Soit $f(v)$ une fonction définie sur chaque sommet v de l'ensemble de données V dans un espace q-dimensionnel et représentée par le tuple : $(f_1, f_2, ..., f_q) \in \mathbb{R}^q$. Ainsi, à chaque sommet $v \in V$ est associé un vecteur de caractéristiques local noté $f(v) \in R^q$. Plusieurs choix peuvent être envisagées pour $f(v)$. Par exemple, pour une séquence d'images, $f(v)$ peut être des attributs spatio-temporels d'un sommet v. Dans le cas le plus simple, nous pouvons considérer que $f(v) = v$. Il existe plusieurs méthodes populaires qui transforment l'ensemble V doté d'une mesure de similarité par paire w en un graphe $G = (V, E, w)$. Parmi les méthodes existantes, nous pouvons citer le **graphe de ε-voisinage** où deux points $u,v \in V$ sont connectés par une arête si $\|f(u) - f(v)\| \leq \varepsilon$, $\varepsilon > 0$. Nous notons alors par $u \sim v$ le fait que le sommet u appartient au ε-voisinage de v ($u \in \mathcal{N}_\varepsilon(v)$) qui est défini par :

$$\mathcal{N}_\varepsilon(v) = \{u \in V,\ \ f(u) = (f'_1, \cdots, f'_q) \ \mid \ |f_i - f'_i| \leq \varepsilon_i, 0 < i \leq q \ \ \} \tag{2.1}$$

Un autre graphe important est le **graphe des k-plus proches voisins** où deux points $u,v \in V$ sont connectés par une arête si u est dans les k-plus proches voisins de v. Le

degré d'un sommet $v \in V$ est la somme des poids des arêtes incidentes à ce sommet :

$$deg(v) = \sum_{u \sim v} w(u,v), \tag{2.2}$$

et le volume du graphe G est :

$$Vol(G) = \sum_{v \in V} deg(v), \tag{2.3}$$

Rappelons que $W = \big(w(u,v)\big)_{(u,v)\in E}$ est la matrice des poids du graphe G et D est une matrice diagonale où toutes les entrées valent 0 sauf celles de la diagonale qui sont égales aux degrés des sommets correspondants, $D = \big(d(v,v) = deg(v)\big)_{v \in V}$.

2.2 Informations et mesures relevées de l'espace observé

L'étape la plus importante dans la définition du graphe $G = (V, E, w)$ est le choix de la fonction des poids w. Cette dernière est étroitement liée à la fonction f qui définit le vecteur des caractéristiques pour chaque objet d'intérêt de la base de données et donc à chaque sommet $v \in V$. Suivant l'application à mettre en place, plusieurs choix des descripteurs ainsi que des mesures de distance sont possibles.

2.2.1 Les descripteurs des objets d'intérêt

La représentation des données multidimensionnelles est un point clé dans toutes les méthodes de classification puisqu'elle permet le passage d'un espace observé à un espace de mesure. Une bonne fonction qui assure ce passage est celle qui minimise la perte d'information entre les deux espaces de représentation.

Nous décrivons dans les sous-sections suivantes, les moments de Zernike 3D, les points SIFT et quelques caractéristiques vidéo que nous avons utilisées dans les applications décrites, respectivement, dans les chapitre 4, 5 et 6.

2.2.1.1 Les moments de Zernike 3D

Pour interpréter le comportement humain dans un clip vidéo, l'objet d'intérêt (OI) doit être identifié et extrait avant sa caractérisation. L'extraction de voxel et en elle-même une tâche difficile ayant ses propres contraintes. Elle n'est donc pas étudiée dans ce manuscrit. Nous supposons que les objets d'intérêt ont déjà été extraits à l'aide d'une

méthode appropriée telle que l'algorithme du "graph cut" [Boykov, 2004, Boykov, 2003], ce qui est utile pour séparer des objets de l'arrière-plan.

Chaque OI est désigné par un sommet $v \in G(V)$ et peut être vu comme un volume binaire exprimé par $g(x, y, t)$. Pour chaque voxel de l'OI, soit g_0 une valeur qui représente sa couleur/intensité initiale et x, y, t ses coordonnées spatio-temporelles.

Pour représenter la façon dont une action est effectuée, nous utilisons les moments de Zernike 3D pour décrire les caractéristiques spatio-temporelles (x, y, t) de l'objet d'intérêt.

Soit $Z^v_{nlm}(x, y, t)$ les fonctions de Zernike 3D :

$$Z^v_{nlm}(x, y, t) = R_{nl}(r) \cdot Y_{lm}(\theta, \phi) \tag{2.4}$$

où $R_{nl}(r)$ est le terme radial, et $Y_{lm}(\theta, \phi)$ sont les harmoniques sphériques du l^{iem} degré, orthonormales sur la surface du sphère unité avec m allant de $-l$ à l et $n-l$ étant un entier non négatif pair $(n-l=2k)$.

L'égalité $(n-l)/2 = k$ est la condition d'orthonormalité des polynômes de Zernike 3D à l'interieur du sphère unité (pour plus de détails, voir [Canterakis, 1999, Novotni, 2004]).

$Y_{lm}(\theta, \phi)$ est le terme angulaire. Z^v_{nlm} peut être réécrit sous une forme plus compacte comme une combinaison linéaire de monômes d'ordres jusqu'à n :

$$Z^v_{nlm}(x, y, t) = \sum_{p+q+r \le n} \mathcal{X}^{pqr}_{nlm} x^p y^q t^r \tag{2.5}$$

où, pour $k = (n-l)/2$:

$$\begin{aligned} \mathcal{X}^{pqr}_{nlm} = & c_{lm} 2^{-m} \sum_{s=0}^{k} q_{kls} \sum_{\alpha=0}^{s} \binom{s}{\alpha} \sum_{\beta=0}^{s-\alpha} \binom{s-\alpha}{\beta} \sum_{r=0}^{m} (-1)^{m-r} \binom{m}{r} \\ & i^r \sum_{\mu=0}^{(l-m)/2} (-1)^{\mu} 2^{-2\mu} \binom{l}{\mu} \binom{l-\mu}{m+\mu} \sum_{s=0}^{\mu} \binom{\mu}{s}, \end{aligned} \tag{2.6}$$

et le facteur de normalisation C_{lm} est donné par :

$$c_{lm} = \frac{\sqrt{(2l+1)(l+m)!(l-m)!}}{l!}, \tag{2.7}$$

et :

$$q_{kls} = \frac{(-1)^k}{2^{2k}} \sqrt{\frac{2l+4k+3}{3}} \binom{2k}{k} (-1)^s \frac{\binom{k}{k}\binom{2(k+l+s)+1)}{2k}}{\binom{k+l+s}{k}} \tag{2.8}$$

Comme Z^{v}_{nlm} forme un système orthonormal complet, il est possible d'approximer la fonction originale g par un nombre fini de moments de Zernike 3D Ω^{v}_{nlm}, comme suit :

$$g(x,y,t) = \sum_{n=0}^{\infty} \sum_{l=0}^{n} \sum_{m=-l}^{l} \Omega^{v}_{nlm} Z^{v}_{nlm}(x,y,t) \tag{2.9}$$

Les moments de Zernike 3D sont définis par :

$$\Omega^{v}_{nlm} = \frac{3}{4\pi} \sum_{p+q+r \leq n} (-1)^m \mathscr{X}^{pqr}_{nlm} m^{v}_{pqr} \tag{2.10}$$

m^{v}_{pqr} désigne les moments géométriques de l'ordre $(p+q+r)$ du volume binaire défini par :

$$m^{v}_{pqr} = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \sum_{t=0}^{N_t-1} x^p y^q t^r g(x,y,t) \tag{2.11}$$

Le choix de l'ordre maximal des moments de Zernike 3D est crucial pour décrire le comportement du sujet et, par conséquent, porter plus ou moins de détails sur le volume binaire vidéo. Ce choix est effectué expérimentalement pour former le vecteur des descripteurs $f(v)$ pour chaque vidéo v. Il est défini par les $(2l+1)$ moments comme suit :

$$f(v) = \{\mathcal{V}^{v}_{nl} = \|\Omega^{v}_{nlm}\| \ : \ n \in [0,N], l \in [0,n], m \in [-l,l]\} \tag{2.12}$$

La distance entre deux vidéos représentées par u et v, respectivement, est calculée en utilisant leurs moment de Zernike 3D comme suit :

$$\big\| f(u) - f(v) \big\| = \|\mathcal{V}^{u}_{nl} - \mathcal{V}^{v}_{nl}\| = \sqrt{\sum_{n=0}^{N} \sum_{l=0}^{n} (\mathcal{V}^{u}_{nl} - \mathcal{V}^{v}_{nl})^2} \tag{2.13}$$

2.2.1.2 Les points SIFT (Scale-invariant feature transform)

Pour décrire une image, une variété de descripteurs de points-clés ont été proposées tels que les détecteurs de coins de Harris [Harris, 1988], la transformation de caractéristiques visuelles invariante à l'échelle (Scale Invariant Feature Transform : SIFT) [Lowe, 2004], l'histogramme du gradient de location et d'orientation (Gradient Location and Orientation Histogram : GLOH) [Mikolajczyk, 2005] et la différence des moyennes (Dom) [Bay, 2008]. Les descripteurs SIFT sont invariants aux changements d'échelle, à la rotation et à l'illumination. En outre, ils sont relativement faciles à extraire et à comparer avec un grand nombre de caractéristiques locales. Plusieurs améliorations des caractéristiques SIFT sont proposées, y compris ASIFT [Morel, 2009] et PCA-SIFT [Ke, 2004] qui utilise l'Analyse en Composantes Principales (ACP) pour réduire la dimensionnalité du descripteur SIFT de 128 à 36.

L'image 2.1, illustre un ensemble de points SIFT, $\mathscr{X} = \{Pt_1, Pt_2, \ldots, Pt_n\}$, extrait en utilisant la méthode décrite dans [Lowe, 2004].

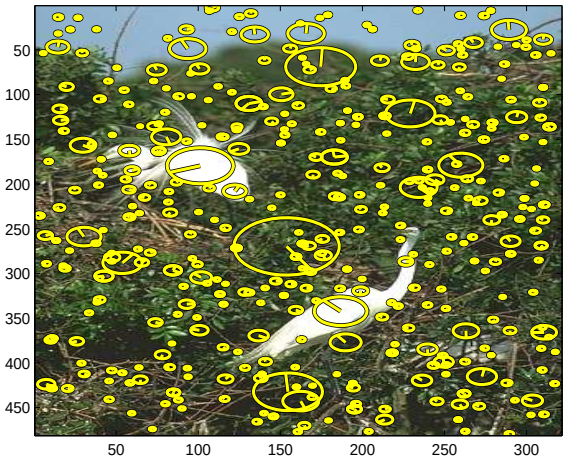

FIGURE 2.1 – Extraction des points SIFT

Chaque point SIFT $Pt_i = (X_i, R_i, U_i)$ est décrit par sa localisation 2D sur l'image $X_i = (x_i, y_i)$, l'amplitude de son gradient et son orientation $R_i = (r_i, a_i)$ et un vecteur descripteur $U_i = (u_{i,1}, \ldots, u_{i,128})$, représentant sa texture locale dans l'image.

2.2.1.3 Des descripteurs pour la vidéo

Les vidéos utilisées dans la partie expérimentation du chapitre 6 sont, d'abord, découpées en images où chacune d'elles est décrite par un vecteur de caractéristiques comportant 7 entrées qui sont :

1. **Les deux couleurs dominantes :** Pour une image donnée, nous retenons les deux premières couleurs dominantes représentées par les deux triplets (h_1, l_1, s_1) et (h_2, l_2, s_2) à partir de l'histogramme de couleur en HLS (hue, luminance, et saturation). (h_1, l_1, s_1) correspond à la luminance l_1 la plus représentée des pixels ayant la saturation s_1 la plus fréquente et cela après avoir extrait la teinte h_1 la plus présente. La deuxième couleur dominante est celle ayant les valeurs h_2, l_2, s_2 les plus représentées tout en étant les plus éloignées, respectivement, de h_1, l_1, s_1. L'objectif de cette procédure est de mettre en exergue le contraste de l'image.
2. **La luminance moyenne :** La variation des valeurs de la luminance moyenne se situe entre 0 et 255.
3. **Le contraste :** Le contraste représente l'éloignement entre les deux couleurs dominantes représentées par les deux triplets (h_1, l_1, s_1) et (h_2, l_2, s_2) .
4. **Les orientations et granularités de texture :** C'est le nombre de pixels voisins présentant de forts changements d'intensité dans chacune des orientations, verticale et horizontale. Cela permet de produire une estimation du taux de haute fréquence verticale et du taux de haute fréquence horizontale.
5. **Le taux d'activité :** C'est le nombre de pixels ayant changé significativement d'intensité entre deux images successives.

2.2.2 Les distances dans l'espace observé

Intuitivement, la distance mesure la longueur (l'intervalle) qui sépare deux objets. Dans un sens plus général, nous pouvons exploiter la notion de distance pour mesurer la différence et ainsi le degré de similarité entre les objets. Dans ce qui suit, nous présentons quelques modèles de distances que nous avons utilisés :

Soit $f(v) = (f_1, \cdots, f_q)$ le vecteur des caractéristiques d'un sommet v, alors :

– **le test du χ^2 :**

$$\|f(u) - f(v)\|^2 = \sum_{i=1}^{q} \frac{(f_i^u - f_i^v)^2}{f_i^u + f_i^v} \tag{2.14}$$

– **Le test de Kolmogorov-Smirnov :**

$$\|f(u) - f(v)\|^2 = \sup_i |S_i^u - S_i^v|, \quad S_i^u = \sum_{j=1}^{i} s(j). \tag{2.15}$$

– **La distance de Bhattacharyya :**

$$\|f(u) - f(v)\|^2 = \left(1 - \sum_{i=1}^{q} \frac{\sqrt{f_i^u f_i^v}}{\sqrt{\Sigma_{i=1}^{q} f_i^u \ \Sigma_{i=1}^{q} f_i^u}}\right)^{1/2} \tag{2.16}$$

– **La distance Intersection :**

$$\|f(u)-f(v)\|^2 = \sum_{i=1}^{q} \min(f_i^u, f_i^v) \tag{2.17}$$

– **La distance Corrélation :**

$$\|f(u)-f(v)\|^2 = \frac{\Sigma_{i=1}^{q} g_i^u g_i^v}{\sqrt{\Sigma_{i=1}^{q}(g_i^u g_i^v)^2}} \tag{2.18}$$

où :

$$g_i^v = \Sigma_{i=1}^{q} f_i^v - \frac{1}{q}\Sigma_{i=1}^{q} f_i^v$$

2.2.3 Mesures de similarité entre les objets d'intérêt

La similarité est une mesure inversement proportionnelle à la distance. Dans le graphe G, la fonction des poids w est utilisée pour mesurer la similarité entre les sommets connectés.

La construction des graphes de similarité consiste à modéliser les relations de voisinage locales et non locales entre les objets de la base de données. Les similarités entre ces objets sont estimées en comparant leurs caractéristiques respectives, ce qui dépend, généralement, de la fonction f et l'ensemble V. Ainsi, définissons un vecteur de caractéristiques non local noté $F(v) \in R^p$ et calculé à partir du patch qui entoure le sommet v comme suit :

$$F(v) = [f(u), u \in \mathscr{B}_k(v) \subseteq \mathscr{N}_\varepsilon(v)]^T \tag{2.19}$$

$\mathscr{B}_k(v)$ est un cadre de sélection de taille k centré sur v qui définit les k-plus proches voisins de v (ainsi, $u \sim v := u \in \mathscr{B}_k(v) \subseteq \mathscr{N}_\varepsilon(v)$). Par conséquent, la fonction de poids w associé au graphe des k-plus proches voisins G, peut incorporer des caractéristiques locales et/ou non locales selon la topologie du graphe considéré. Elle donne une mesure de la similarité entre un sommet et ses voisins qui peut incorporer des caractéristiques locales et non locales. Elle est définie comme suit :

$$w(u,v) = \begin{cases} exp\left(-\frac{\|f(u)-f(v)\|^2}{h_1^2}\right).exp\left(-\frac{\|F(u)-F(v)\|^2}{h_2^2}\right), \text{ si } u \sim v \\ 0 \quad \text{sinon} \end{cases} \tag{2.20}$$

le paramètre d'échelle h_i peut être estimé en utilisant l'écart type en fonction de la variation, respectivement, de $\|f(u)-f(v)\|$ et de $\|F(u)-F(v)\|$ sur le graphe.

2.3 Marches aléatoires sur le graphe

Le graphe G reflète la connaissance de la géométrie locale / non locale des objets d'intérêt de la base de données. La normalisation de la matrice W telle que :

$$P = D^{-1}W \tag{2.21}$$

permet de définir une matrice stochastique P ($0 \leq p(u,v) \leq 1$, car $0 \leq w(u,v) \leq deg(v)$). La matrice P est directement liée à la matrice des poids W. Alors, P permet de capturer, aussi, les relations mutuelles entre les objets d'intérêt.

P peut être interprétée comme étant la matrice de transition d'une chaîne de Markov dont les états de cette dernière sont, dans notre cas, les sommets du graphe. Ainsi, une marche aléatoire sur le graphe G est le processus qui part d'un sommet et, à chaque instant, se déplace vers un autre sommet avec une probabilité proportionnelle au poids de l'arête correspondante. Ainsi, nous pouvons définir la diffusion sur G comme étant l'ensemble des éventuels sommets visités à partir d'un sommet donné, où une transition est effectuée en *une seule étape* du sommet u vers un autre sommet v choisis aléatoirement et uniformément parmi son voisinage, avec la probabilité :

$$\begin{aligned} p^{(1)}(u,v) &= Pr(X_{t+1} = v | X_t = u) \\ &= \frac{w(u,v)}{deg(u)} \end{aligned} \tag{2.22}$$

La matrice de transition P correspondante à G est donnée par : $P = \{p^{(1)}(u,v) \mid u,v \in V,\ u \sim v\}$. Elle explicite toute les transitions possibles en *une seule étape* et fournit donc les informations du premier ordre sur la structure de ce graphe.

soit P^t la t^{ieme} puissance de la matrice P qui représente l'ensemble de toutes les probabilités de transition $p^{(t)}(u,v)$ de partir d'un sommet à un autre en t-$étapes$. Donc, sur le graphe G, $p^{(t)}(u,v)$ reflète tous les chemins de longueur t entre un sommet u et un sommet v. Ainsi, la matrice P^t permet de considérer des voisinages plus étendus. Ces probabilités de transition en t-$étapes$ satisfont l'équation de Chapman–Kolmogorov, où pour tout k tel que $0 < k < t$:

$$\begin{aligned} p^{(t)}(u,v) &= Pr(X_t = v | X_0 = u) \\ &= \sum_{y \in V} p^{(k)}(u,y) \cdot p^{(t-k)}(y,v) \end{aligned} \tag{2.23}$$

Le graphe G étant connexe et chacun de ces sommets a k voisins (G est un graphe des k- plus proche voisins). Par conséquent, la chaîne de Markov associée est réversible et les marches aléatoires convergent vers une *distribution stationnaire unique* $\pi = (\pi_1, \cdots, \pi_n)$ telle que :

$$P^t \pi = \pi \tag{2.24}$$

π a une double interprétations. Elle correspond d'une part, aux taux d'occupation des états de la chaîne de Markov et par conséquent, des sommets du graphe G, et d'autre part, aux probabilité de visites des chacun des sommets de G. Par exemple, π_v est la probabilité d'être au sommet v à partir de n'importe quel autre sommet $u \in V$. Elle est alors proportionnelle à son poids [Lafon, 2006] :

$$\pi_v = \frac{deg(v)}{Vol(G)} = \lim_{t \to +\infty} p^t(u, v) \tag{2.25}$$

2.3.1 Cartes de diffusion

$p(u, v)$ représente la probabilité de transition en un seul saut (un chemin de longueur 1) d'un sommet u à un autre sommet v. Ainsi, la matrice P reflète la géométrie locale définie par le voisinage direct (immédiat) de chaque sommet du graphe de données. Le processus de diffusion sur le graphe G, matérialisé par les puissances de la matrice P permet, par contre, de révéler les structures géométriques pertinentes de la base de données à différentes échelles (à différents instants t). De gauche vers la droite, la figure 2.2 montre la structure de la matrice P à différentes échelles t.

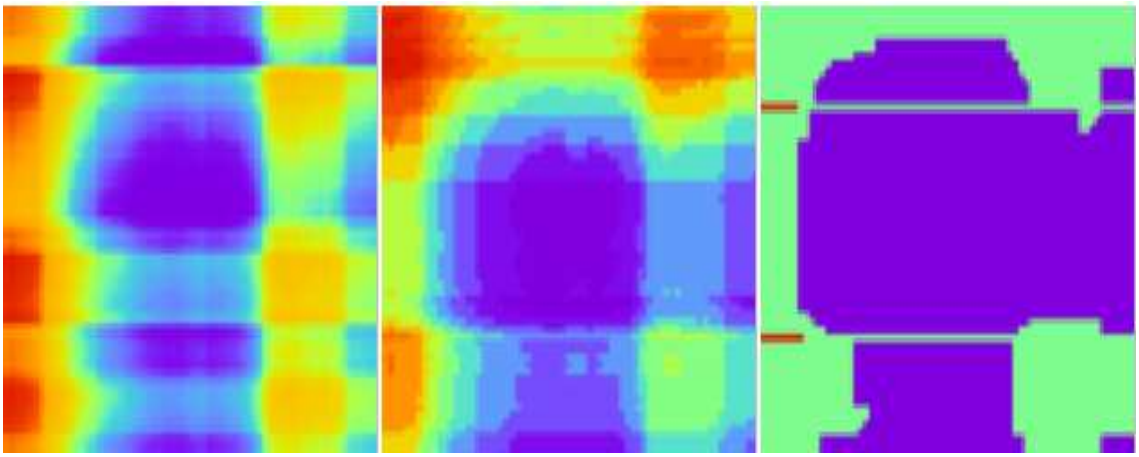

FIGURE 2.2 – La structure de la matrice P à différents instants t

Par conséquent, l'étude des propriétés spectrales de P^t permet d'étudier la géométrie de l'ensemble des objets d'intérêt et de déduire une représentation compacte de l'es-

pace observé en définissant de nouvelles coordonnées composées principalement des valeurs et vecteurs propres significatifs de la matrice P. Ce qui constitue une réduction non linéaire de la dimensionnalité.

La transformation suivante :

$$S = D^{1/2} P D^{-1/2} \tag{2.26}$$

permet de définir la matrice symétrique S qui est similaire à la matrice P. Ainsi, P et S partagent les mêmes valeurs propres. De plus, étant symétrique et à éléments réels, S est diagonalisable à l'aide d'une matrice orthogonale. Ses valeurs propres, $\{\lambda_i\}_{i=0}^{n-1}$, sont donc réelles et ses sous-espaces propres sont orthogonaux (les vecteurs propres correspondant $\{v_i\}$ forment une base orthonormale dans $\mathbb{R}^n$). De plus, la matrice P étant stochastique et le graphe G étant connexe, alors P (et S), possède une seule valeur propre dominante $\lambda_0 = 1$, les valeurs restantes sont ordonnées par la relation suivante [Nadler, 2005] :

$$1 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0 \tag{2.27}$$

Il est à noter que les valeurs propres de P^t sont $\{\lambda_i^t\}_{i=0}^{n-1}$, et que les vecteurs propres de gauche et de droite de P notés, respectivement, ϕ_i et ψ_i sont reliés à ceux de S par :

$$\phi_i = v_i D^{1/2}, \quad \psi_i = v_i D^{-1/2} \tag{2.28}$$

Ainsi, les vecteurs ϕ_i et ψ_i sont aussi linéairement indépendants et nous pouvons écrire :

$$\langle \phi_i, \psi_j \rangle = \delta_{i,j} \tag{2.29}$$

$\delta_{i,j}$ est le symbole de Kronecker, où $\delta_{i,j} = 0$ si $i \neq j$ et $\delta_{i,j} = 1$ si $i = j$.

Les vecteurs propres gauches de la matrice P vérifient la relation suivante :

$$\phi_i P = \lambda_i \phi_i \tag{2.30}$$

Ainsi, pour $\lambda_0 = 1$, nous obtenons :

$$\phi_0 P = \phi_0 \tag{2.31}$$

La matrice P est une matrice stochastique irréductible (G est un graphe k-pp connexe sur un ensemble V fini de données et $p(u,v) \geq 0$). Alors, P admet *une unique probabilité invariante* π. Donc nous pouvons écrire aussi :

$$\pi P = \pi \tag{2.32}$$

Ainsi, pour un durée t suffisamment grande pour que les marche aléatoires convergent vers la distribution stationnaire définie en 2.24, nous déduisons que :

$$\phi_0 = \pi \tag{2.33}$$

Ainsi,

$$\lim_{t \to +\infty} p^t(u,v) = \pi_v = \frac{deg(v)}{Vol(G)} = \phi_0(v) \tag{2.34}$$

et ce, indépendamment de la position de départ $u \in V$.

Pour tout autre instant fini t, la probabilité $p^t(u,v)$ est décomposée dans la base orthonormale des vecteurs propres $\{\phi_i\}_{i=1}^{n-1}$ en [Nadler, 2005] :

$$p^t(u,v) = \phi_0(v) + \sum_{i \geq 1} a_i(u) \lambda_i^t \phi_i(v) \tag{2.35}$$

où le coefficient a_i dépend de la position initiale u.

En utilisant la condition d'orthogonalité définie en 2.29, nous pouvons écrire : $a_i(u) = \psi_i(u)$, avec $a_0(u) = \psi_0(u) = 1$.

2.3.2 Les nouvelles coordonnées dans l'espace réduit

Par définition, les marches aléatoires sur le graphe G peuvent atteindre un régime stationnaire avec une distribution d'occupation unique π des états de la chaîne de Markov et ainsi des sommets du graphe. Par conséquent, nous pouvons considérer que deux sommets u, v sont similaires s'ils ont des distributions π_u et π_v égales (le cas idéal) ou très proche l'une de l'autre. Ainsi, nous pouvons définir une famille de distances sur le graphe G basées sur les probabilités de passage entre sommets et ce à différentes échelles t.

La **distance de diffusion** [Nadler, 2005] sur le graphe est donc :

$$\begin{aligned} D_t^2(u,v) &= \| p^t(u,z) - p^t(v,z) \|_g \\ &= \sum_{z \in V} \left(p^t(u,z) - p^t(v,z) \right)^2 g(z) \end{aligned} \tag{2.36}$$

Le choix $g(z) = 1/\phi_0(z)$, permet de prendre en compte la densité locale des points de la base de données (ϕ_0 étant la distribution d'occupation des états de la chaîne de Markov associée au graphe de données G).

Sur un autre plan, les valeurs et vecteurs propres de la matrice P,$\{\lambda_i^t,\ \psi^i(u)\}$, permettent de générer des coordonnées euclidiennes pour la représentation des sommets du graphe G dans l'espace réduit à différents instants t où pour chaque sommet, ces coordonnées sont données par :

$$\Psi_t(v) = \left(\lambda_1^t \psi^1(v), \lambda_2^t \psi^2(v), \ldots, \lambda_n^t \psi^n(v) \right)^T \tag{2.37}$$

Elles correspondent à la bijection non-linéaire des sommets du graphe dans le nouvel espace réduit euclidien.

Une nouvelle distance **Euclidienne**, et égale à 2.36, est alors définie en utilisant ces nouvelle coordonnées. Cette nouvelle distance appelée **Diffusion maps** [Coifman, 2006], est donnée par :

$$\begin{aligned} D_t^2(u,v) &= \sum_{i \geq 1} \lambda_i^{t2} (\psi_i(u) - \psi_i(v))^2 \\ &= \left\| \Psi_t(u) - \Psi_t(v) \right\|^2 \end{aligned} \tag{2.38}$$

2.3.3 L'information utile dans l'espace réduit

Les valeurs propres, ainsi extraites, sont uniques et ordonnées telles que : $1 = |\lambda_1| \geq |\lambda_2| \geq \ldots |\lambda_n| \geq 0$. Par conséquent, les premières valeurs propres dominantes portent l'information utile sur les objets d'intérêt et sont bien adaptées pour définir les nouvelles coordonnées euclidiennes. Pratiquement, nous pouvons utiliser le test subjectif de Cattell [Cattell, 1966] pour déterminer les k^{ieme} plus importantes dimensions qui capturent l'information pertinente. Ce critère est basé sur l'analyse des différences entre les valeurs propres consécutives, où un point de rupture se trouve là où il y a le plus grand changement dans la pente de la courbe des valeurs propres. Les premiers k^{ieme} valeurs propres correspondent donc au nombre de dimensions à prendre en compte. Une autre méthode simple consiste à considérer les premières valeurs propres dominantes dont la somme est supérieure à un seuil défini (par exemple. $\geq 80\%$).

A titre illustratif, pour reconnaître des comportements spécifiques dans des clips vidéos (cf. chapitre suivant) initialement observés comme dans la figure 2.3, il est visiblement clair que la topologie des données est non linéaire. De plus, l'exécution d'un algorithme de classification directement sur ce nuage génère des erreurs de classification qui induisent des ambiguïtés dans la classification des actions (une couleur par action), en particulier entre *jogging, walking, running and boxing.*

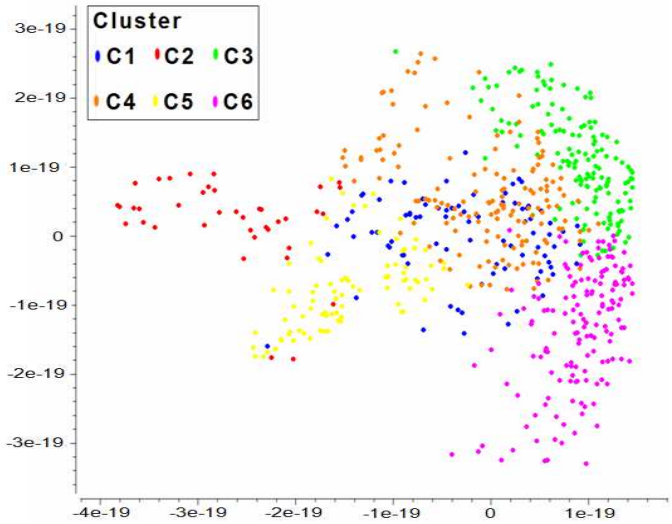

FIGURE 2.3 – Les actions de la base KTH projetées dans l'espace des caractéristiques

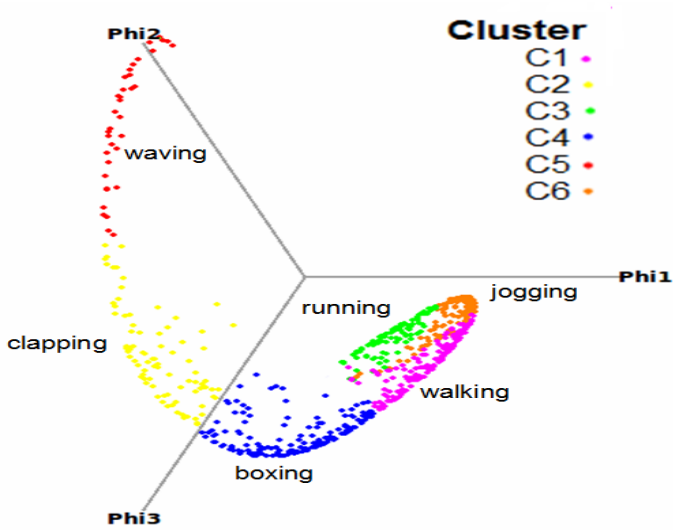

FIGURE 2.4 – Les actions de la base KTH projetées dans l'espace réduit

En revanche, la figure 2.4 montre la projection des mêmes vidéos dans l'espace réduit. Chaque vidéo est représentée par les 10 premières coordonnées. Ici, les classes des vidéos sont correctement séparées et la catégorisation de la base des vidéos est clairement visible.

Il convient de mentionner que pour les base de données de grande dimension, nous pouvons utiliser des solveurs itératifs pour déterminer une partie des valeurs propres.

2.3.4 Accélération des marches sur le graphe

Les valeurs propres de la matrice P correspondent a la solution de son équation caractéristique :

$$\lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \ldots + c_0 = 0 \tag{2.39}$$

Pour un ensemble de données de grande dimension (grandes valeurs de n), la résolution de cette équation prend beaucoup de temps et ralentit ainsi le calcul des valeurs et vecteurs propres.

De nombreux travaux dans la littérature, y compris la théorie des perturbations des matrices [Stewart, 1990, Li, 2003] et la méthode de Nyström [Nyström, 1930, Baker, 1977], ont été proposés pour accélérer la décomposition spectrale par l'approximation des valeurs propres et les vecteurs propres correspondants. Dans [Antonio, 2002], l'approximation d'un vecteur propre dominant est basée sur une analyse de perturbation linéaire des matrices non négatives et symétriques. L. Huangn et al. ont étudié dans [Huang, 2008] les effets de la perturbation des données sur la performance du regroupement spectral et sa relation avec la perturbation des valeurs et vecteurs propres de la matrice Laplacienne. A son tour, la méthode de Nyström a eu beaucoup de succès. Dans [Williams, 2001], les auteurs montrent son utilisation pour approximer le spectre de la matrice de Gram pour accélérer les machines à noyaux.

Une méthode alternative pour approximer les valeurs et vecteurs propres, est l'utilisation de *l'algorithme des puissances itérées* et *la méthode de déflation* pour chercher le vecteur propre dominant et la valeur propre correspondante. Cette méthode exploite le fait que la décomposition de la matrice P fournit un ensemble de valeurs propres ordonnées comme dans 2.27.

En effet, l'algorithme des puissances itérées est une méthode simple pour calculer le vecteur propre dominant car il accède à la matrice uniquement à travers sa multiplication par des vecteurs. Cette propriété est particulièrement importante dans le cas des matrices de très haut rang. La méthode de déflation, quant à elle, permet de retirer à chaque itération, la valeur propre dominante et réarrange la matrice de telle sorte que la valeur propre dominante de cette nouvelle matrice n'est autre que la deuxième valeur propre de la matrice d'origine. Ce processus est répété pour calculer les valeurs propres une par une.

2.3.4.1 L'algorithme des puissances itérées

Pour avoir une bonne approximation du vecteur propre dominant de la matrice P, nous pouvons choisir une approximation initiale V^0 qui doit être un vecteur non nul dans $\mathbb{R}^n$ de telle sorte que sa multiplication par P converge vers le vecteur propre dominant. L'algorithme suivant résume la méthode des puissances itérées.

Algorithme 1 : L'algorithme des puissances itérées

ENTRÉES : V^0, un vecteur non nul dans $\mathbb{R}^n$
SORTIES : Une approximation du vecteur propre dominant
1: **tantque** $\| V^k - V^{k-1} \| / \| V^k \| \geq \varepsilon$ **faire**
2: Poser $X^k = PV^{k-1}$
3: Poser α^k = le plus grand élément de X^k (en valeur absolue)
4: Poser $V^k = X^k / \alpha^k$
5: **fin tantque**
6: **return** V^k, une approximation du vecteur propre dominant de P

Diviser par α^k dans la ligne 4, consiste à réduire chaque approximation avant de passer à l'itération suivante afin d'éviter des vecteurs dont les composantes sont trop grandes (ou trop petites). Pour des grandes puissance, k, nos obtenons une bonne approximation du vecteur propre dominant. En effet, P possède une base orthonormale de vecteurs propres $\{\psi_i\}$. Ainsi, l'approximation initiale V^0 peut être écrite sous la forme :

$$V^0 = \sum_{i=1}^{n} \beta_i \psi_i \ , \quad \beta_i \in \mathbb{R} \tag{2.40}$$

Supposons que ψ_1 est le vecteur propre correspondant à la valeur dominante λ_1, alors nous pouvons facilement écrire :

$$\begin{aligned} V^k = P^k V^0 &= \sum_{i=1}^{n} \beta_i P^k \psi_i = \sum_{i=1}^{n} \beta_i \lambda_i^k \psi_i \\ &= \beta_1 \lambda_1^k \left\{ \psi_1 + \ldots + \sum_{i=2}^{n} \frac{\beta_i}{\beta_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \psi_i \right\} \end{aligned} \tag{2.41}$$

Comme λ_1 est la valeur propre dominante, il s'ensuit que :

$\lambda_i / \lambda_1 < 1$, et $\forall i > 1$, $\lim_{k \to \infty} (\lambda_i / \lambda_1)^k \longrightarrow 0$. Nous en déduisons alors que $P_t^k V^0 \approx \beta_1 \lambda_1^k \psi_1$, $\beta_1 \neq 0$. Ce qui signifie que la direction de V^k se stabilise à celle de ψ_1 et comme ψ_1

est un vecteur propre dominant, il s'ensuit que tout scalaire multiple de ψ_1 est aussi un vecteur propre dominant. En outre, puisque les valeurs propres de P sont ordonnées comme dans 2.27, alors la méthode des puissances itérées converge rapidement si $|\lambda_1|/|\lambda_2|$ est petit, et lentement si $|\lambda_1|/|\lambda_2|$ est très proche de 1.

2.3.4.2 La méthode de déflation

Une fois une approximation du vecteur propre dominant ψ_i est calculée, *le quotient de Rayleigh* permet de fournir une bonne approximation à la valeur propre dominante correspondante λ_i qui est donnée par :

$$\lambda_i = \frac{(P_i\psi_i)^T \cdot \psi_i}{\psi_i^T \psi_i} \tag{2.42}$$

Pour calculer les valeurs propres restantes, la matrice P_i est modifié en $P_{i+1}, \ldots,$ comme suit :

$$P_{i+1} = P_i - \lambda_i \frac{\psi_i \psi_i^T}{\psi_i^T \psi_i} \tag{2.43}$$

P_{i+1} a les même vecteurs et valeurs propres que P_i à l'exception de λ_i qui est décalée à 0 laissant les autres valeurs propres inchangées. En effet, pour tout vecteur propre $\psi_j,\ j = (i+1, i+2, \cdots, n)$, de P, P_{i+1} satisfait :

$$\begin{aligned} P_{i+1}^t \psi_j &= P_i^t \psi_j - \lambda_i \frac{(\psi_i \psi_i^T) \cdot \psi_j}{\psi_i^T \psi_i} \\ &= P^t \psi_j - \lambda_i \frac{\psi_i \cdot (\psi_i^T \psi_j)}{\psi_i^T \psi_i} \end{aligned} \tag{2.44}$$

Comme l'ensemble des vecteurs propres $\{\psi_i\}$ forme une base orthonormale (c.à.d : $\psi_i^T \psi_j = 0$) , alors : $P_{i+1}\psi_j = P_i\psi_j$. Ainsi, les vecteurs propres de P_{i+1} sont les mêmes que ceux de P_i et ses valeurs propres sont λ_{i+1}, $\cdots$,λ_n. L'algorithme des puissances itérées appliqué à P_{i+1} va trouver, alors, la prochaine grande valeur propre λ_{i+1}.

Pour vérifier cette méthode, nous avons effectué des tests sur différentes matrices de tailles différentes. Nous étions limités aux 10^{ieme} premiers vecteurs et valeurs propres importants. Par exemple, dans la figure 2.5, l'écart entre les valeurs propres (eigengap) est bien observé entre la première et la deuxième valeur propre.

Comme le montre la figure 2.6, il n'y a pratiquement pas de différence entre les valeurs propres obtenues en utilisant cette technique et celles calculées en utilisent la

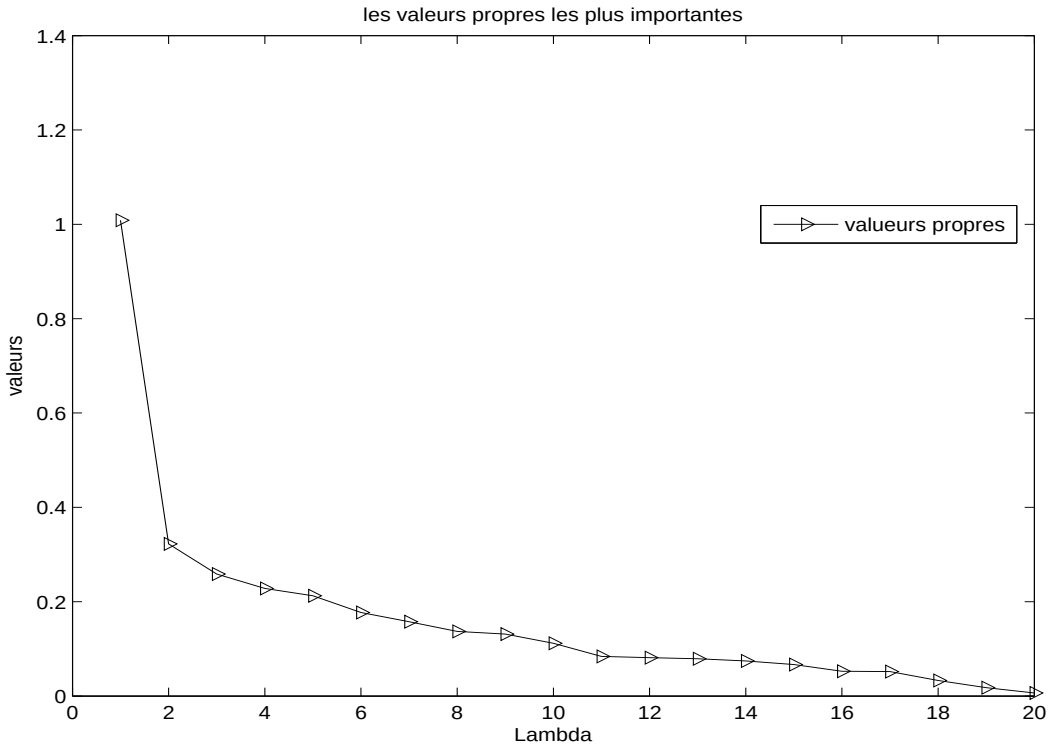

FIGURE 2.5 – Les valeurs propres les plus importantes

méthode SVD (Singular Value Decomposition). Cependant, concernant le temps de calcul (voir la figure 2.7), il est clair que l'approche des déflations itératives est plus efficace car elle ne calcule que les premiers valeurs propres tandis que la méthode SVD doit décomposer la matrice toute entière pour extraire les valeurs et vecteurs propres considérés et par conséquent, nécessite plus de temps.

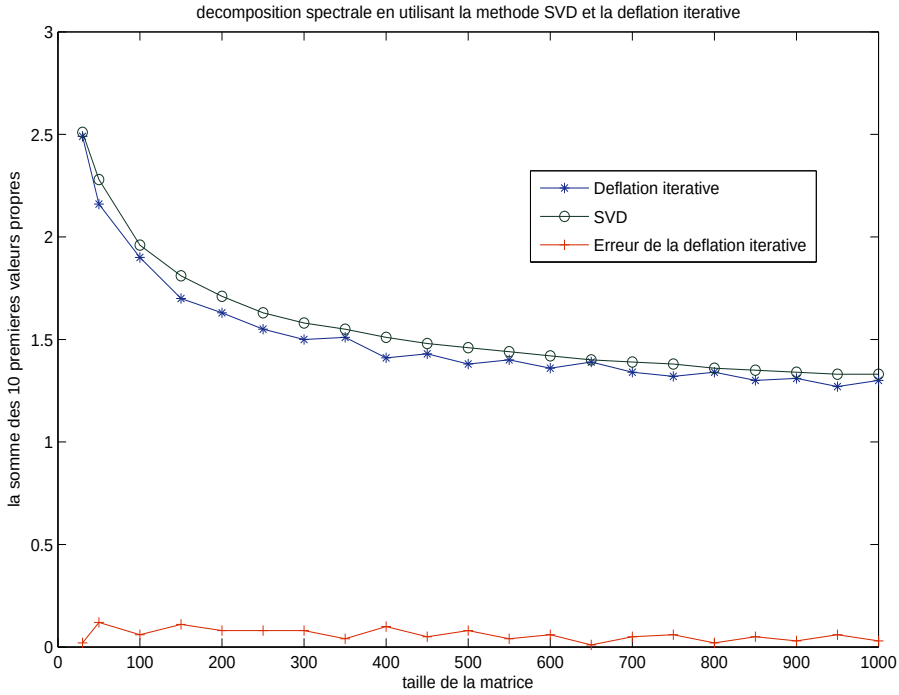

FIGURE 2.6 – La somme des 10 premières valeurs propres en utilisant la méthode SVD et les déflations itératives

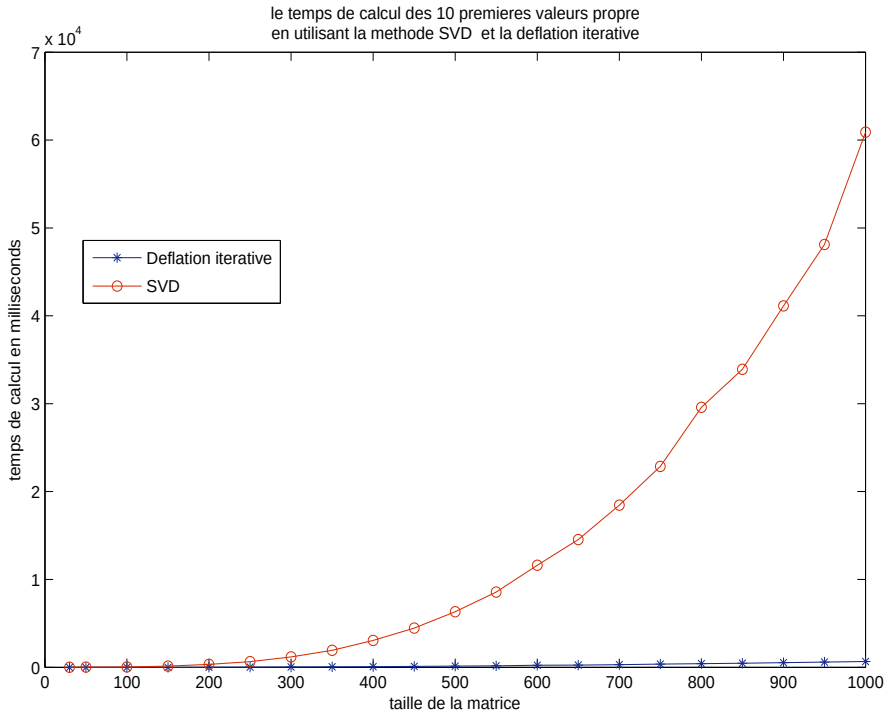

FIGURE 2.7 – Le temps de calcul des 10 premières valeurs en utilisant la méthode SVD et les déflations itératives

Chapitre 3

Régularisation discrète sur graphes

Sommaire

3.1 Introduction

Si la réduction des données permet d'améliorer la lisibilité des données en vue de leur analyse, d'autres approches exploitent les informations fournies par un sous ensemble de données et diffusent les connaissances acquises sur le reste de l'ensemble afin de le classifier. Parmi ces approches, nous avons la régularisation discrète sur graphe où le processus de diffusion se base sur le concept de l'inférence transductive et les interactions entre les données sont facilitées par un graphe capturant les relations mutuelles entre elles. Ceci permet ainsi, la diffusion progressive des connaissances.

Les sections suivantes décrivent quelques notions et définitions sur les graphes ainsi que les opérateurs élémentaires qui interviennent dans la description des méthodes de régularisation.

3.2 Fonctions et opérateurs sur graphes

Le degré d'un sommet $v \in V$ est la somme des poids des arêtes incidentes à ce sommet :

$$deg(v) = \sum_{u \sim v} w(u, v), \tag{3.1}$$

et le volume du graphe G est :

$$Vol(G) = \sum_{v \in V} deg(v), \tag{3.2}$$

Rappelons que $W = \big(w(u,v)\big)_{(u,v)\in E}$ est la matrice des poids du graphe G et D est une matrice diagonale où toutes les entrées valent 0 sauf celles de la diagonale qui sont égales aux degrés des sommets correspondants, $D = \big(d(v,v) = deg(v)\big)_{v\in V}$.

3.2.1 Opérateurs différence, dérivée partielle et adjoint

Soit $f : V \to \mathbb{R}$ une fonction qui attribue à chaque sommet $v \in V$ une valeur réelle $f(v)$ et soit $\mathscr{H}(V)$ l'espace de Hilbert des fonctions réelles muni du produit interne :

$$\langle f, g \rangle_{\mathscr{H}(V)} = \sum_{V} f(v)g(v), \tag{3.3}$$

et de la norme $\mathscr{L}_2$, où $\|f\|_2 = \langle f, f \rangle^{1/2}_{\mathscr{H}(V)}$.

De la même manière, soit $\mathscr{H}(E)$ l'espace de Hilbert des fonctions à valeurs réelles, $F : E \to \mathbb{R}$, définies sur les arêtes du graphe G muni du produit interne :

$$\langle F, H \rangle_{\mathscr{H}(E)} = \sum_{(u,v)\in E} F(u,v)H(u,v), \tag{3.4}$$

et de la norme $\mathscr{L}_2$, où $\|F\|_2 = \langle F, F \rangle^{1/2}_{\mathscr{H}(E)}$.

L'opérateur *différence pondérée* d'une fonction $f : V \to \mathbb{R}$ en une arête $(u,v) \in E$ est défini par :

$$(d_w f)(u,v) = \frac{f(v) - f(u)}{distance(u,v)} \tag{3.5}$$

$distance(u,v)$ peut prendre plusieurs formes, par exemple : $distance(u,v) = 1/\sqrt{w(u,v)}$. Donc l'opérateur *différence pondérée* s'écrit, alors :

$$(d_w f)(u,v) = \sqrt{w(u,v)}(f(v) - f(u)) \tag{3.6}$$

L'opérateur *dérivée partielle* d'une fonction $f : V \to \mathbb{R}$ par rapport à une arête $e = (u,v) \in E$ en un sommet $v \in V$ est défini par :

$$\left.\frac{\partial f}{\partial e}\right|_v = \partial_u f(v) = (d_w f)(u,v) \tag{3.7}$$

De la définition (3.7), on peut déduire les remarques suivantes :

1. si $f(v) = f(u) \to \partial_u f(v) = 0$. (la dérivée d'une fonction constante est nulle)
2. $\partial_u f(v) = -\partial_v f(u)$. (la fonction w est symétrique et le graphe est non-orienté)
3. $\partial_v f(v) = 0$.

L'opérateur *adjoint* $d^*_v : \mathscr{H}(E) \to \mathscr{H}(V)$ d'une fonction $H \in \mathscr{H}(E)$ s'exprime localement en un sommet $v \in V$ par :

$$(d^*_w H)(v) = \sum_{u\sim v} \sqrt{w(u,v)}(H(u,v) - H(v,u)). \tag{3.8}$$

3.2.2 Opérateurs divergence et gradient

L'opérateur de *divergence*, $div : \mathscr{H}(E) \to \mathscr{H}(V)$, de la fonction $H \in \mathscr{H}(E)$ est défini par : $divH = -d^* H$.

et donc, en un sommet v, par :

$$(-div H)(v) = \sum_{u \sim v} \sqrt{w(u,v)}(H(u,v) - H(v,u)). \tag{3.9}$$

L'opérateur *gradient pondéré* $\nabla_w : \mathscr{H}(V) \rightarrow \mathbb{R}^m$, d'une fonction $f : V \rightarrow \mathbb{R}$ en un sommet $v \in V$, mesure la régularité de la fonction dans le voisinage du sommet. Plusieurs normes ont été proposées. Soit l'expression de la norme $\mathscr{L}_p$ pour $p \in]0, +\infty[$:

$$\|\nabla_w f(v)\|_p = \left(\sum_{u \sim v} \left|\partial_u f(v)\right|^p \right)^{1/p} = \left(\sum_{u \sim v} \left(\sqrt{w(u,v)} \left| f(v) - f(u) \right| \right)^p \right)^{1/p} \tag{3.10}$$

En particulier, pour $p = 2$ nous obtenons la norme $\mathscr{L}_2$, qui se réécrit :

$$\|\nabla_w f(v)\|_2 = \sqrt{\sum_{u \sim v} \left(\partial_u f(v)\right)^2} = \sqrt{\sum_{u \sim v} w(u,v) \Big(f(v) - f(u) \Big)^2} \tag{3.11}$$

3.2.3 Opérateur p-Laplacien

L'opérateur *Laplacien* est utilisé dans de nombreux traitements, en particulier dans des processus de diffusion. Il est directement lié à la théorie spectrale des graphes où plusieurs formulations du Laplacien sont décrites en fonction de la méthode d'analyse de données adoptée (Isomap, LLE, etc.). Une des formes possible du *Laplacien* est la suivante :

$$L = D - W \tag{3.12}$$

L'opérateur *Laplacien*, $\triangle_w : \mathscr{H}(V) \rightarrow \mathscr{H}(V)$, d'une fonction $f : V \rightarrow \mathbb{R}$, est défini par :

$$\triangle_w f = \frac{1}{2} d_w^* (d_w f), \tag{3.13}$$

et s'exprime en un sommet $v \in V$ par :

$$\triangle_w f(v) = \sum_{u \sim v} w(u,v) \Big(f(v) - f(u) \Big) \tag{3.14}$$

Une famille des *p-Laplacien* qui inclue l'opérateur donné par (3.14) est la suivante :

$$\triangle_w^p f(v) = \frac{1}{2} \sum_{u \sim v} w(u,v) \Big(\left\| \nabla_w f(v) \right\|_2^{p-2} + \left\| \nabla_w f(u) \right\|_2^{p-2} \Big) \Big(f(v) - f(u) \Big) \tag{3.15}$$

Pour $p = 2$, nous retrouvons le *Laplacien* du second ordre défini en (3.14).

3.3 Cadre général de la régularisation discrète

Dans cette section, nous allons décrire une méthode variationnelle sur graphe (modélisant un espace observé discret et fini) et cela dans le but d'estimer une fonction f, définie sur les sommets de ce graphe, mais dont le modèle est inconnu. Les sommets du graphe représentent des objets (pixels, images, vidéos, page web, etc.) dont les vecteurs des caractéristiques sont donnés par une fonction f^0 décrivant l'observation initiale de ces objets.

Il y a plusieurs situations du monde réel où f est inconnue. Par exemple dans le cas d'une image bruitée, f^0 (l'image bruitée) peut être considérée comme une observation d'une fonction originale f (l'image d'origine) altérée par un bruit η : $f^0 = f + \eta$. Dans ce cas, les sommets du graphe représentent les pixels, les arêtes lient un pixel donné avec ces 4 ou ces 8 voisins. La fonction f en un pixel donnée peut représenter son intensité et/ou couleur et/ou $\cdots$ etc.

Dans d'autres cas, nous disposons d'une connaissance partielle de f, puisque définie sur un sous-ensemble des sommets du graphe, et nous souhaitons estimer la valeur de f sur le reste des sommets.

Ayant un ensemble d'objets $\mathscr{X} = \{x_1, x_2, \cdots, x_l, x_{l+1}, \cdots, x_n\}$, $x_i \in \mathbb{R}^d$, dont les l premiers objets sont labellisés par $y_1, y_2, \cdots, y_l \in \mathscr{Y} = \{1, 2, \cdots, c\}$, le but est de prédire les labels des objets, allant de $l+1, \cdots, n$, non encore labellisés en utilisant une fonction f qui associe à un objet $x_i \in \mathscr{X}$, un label $y_i \in \mathscr{Y}$. Donc, $f^0(x_i) \in \mathscr{Y}$ si x_i est labellisé, et $f^0(x_i) = 0$ sinon.

Définir une fonction f sur la totalité de l'ensemble $\mathscr{X}$ est plus complexe que de la définir directement sur un sous-ensemble de $\mathscr{X}$ pour prédire les classes des objets non encore labellisés à partir des classes de ceux qui le sont déjà. L'idée sous-jacente, à l'estimation de f, repose sur les liaisons mutuelles capturées par le graphe modélisant l'ensemble $\mathscr{X}$. Ainsi, un objet particulier absorbe une partie des connaissances contenues dans son voisinage sans aucune considération du reste de la topologie de l'ensemble $\mathscr{X}$. Par conséquent, nous considérons ici l'estimation d'une fonction de classification discrète qui minimise une certaine perte (due à la non considération de la topologie globale) plus un terme de régularisation (d'ajustement) pour harmoniser la valeur de f sur un objet, avec les valeurs (les classes) du voisinage de cet objet.

La régularisation discrète de la fonction $f^0 \in \mathscr{H}(V)$ en utilisant l'opérateur *Laplacien pondéré* consiste à chercher une fonction $f^* \in \mathscr{H}(V)$ suffisamment régulière sur G, tout en étant suffisamment proche de la fonction f^0. Une approche classique, pour résoudre ce problème, utilise la méthode de régularisation et consiste à minimiser la fonctionnelle à deux termes :

$$\mathscr{E}(f) = E_{lissage}(f) + \lambda E_{attache}(f^0, f) \tag{3.16}$$

$E_{lissage}$ est un terme de régularisation représentant l'information a priori sur la fonction f. $E_{attache}$ est un terme de fidélité aux données.

Les modèles variationnels de régularisation peuvent être décrits par le problème de minimisation suivant :

$$f^* = \min_{f \in \mathscr{H}(V)} \left\{ \frac{1}{2} \sum_{v \in V} \| \nabla f_v \|_2^2 + \frac{\lambda}{2} \| f - f^0 \|^2_{\mathscr{H}(V)} \right\} \tag{3.17}$$

Le paramètre de fidélité λ, appelé aussi multiplicateur de Lagrange, permet d'établir un équilibre entre les deux énergies spécifiées dans (3.17), correspondant respectivement, aux premier terme de lissage (ou régularisateur) et le deuxième terme d'ajustement par rapport aux observations initiales. La solution de ce problème de régularisation peut être obtenue en utilisant l'algorithme itératif de Gauss-Jacobi suivant, où $\forall (u, v) \in E$:

$$\begin{cases} f^{(0)} = f^0 \\ f^{t+1}(v) = \dfrac{1}{\lambda + \sum\limits_{u \sim v} w(u, v)} (\lambda f^0(v) + \sum\limits_{u \sim v} w(u, v) f(u)^t) \end{cases} \tag{3.18}$$

La nouvelle valeur $f^{(t+1)}(v)$ dépend de la valeur initiale (observée) $f^0(v)$ et d'une moyenne pondérée des valeurs existantes dans le voisinage de v. Pour vérifier la convergence de (3.18), il suffit de tester à chaque itération la variation de la fonction f par rapport à l'itération précédente moyennant le taux suivant :

$$\left\| f^{(t+1)} - f^{(t)} \right\|_2 < \varepsilon \left\| f^{(t+1)} \right\|_2 \ , \quad \varepsilon \longrightarrow 0 \tag{3.19}$$

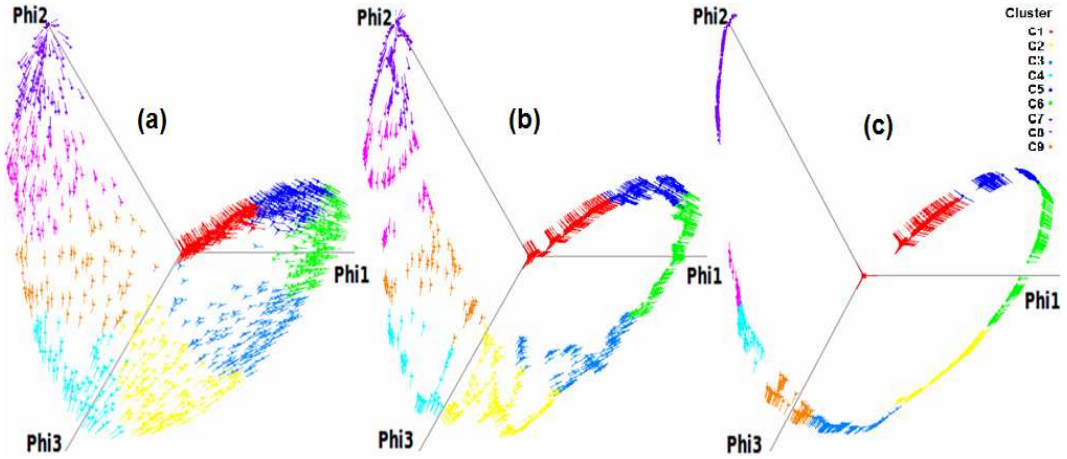

FIGURE 3.1 – Régularisation discrète avec différents p

La figure 3.1-a représente la projection d'un nuage de points (des objets observés) dans un espace tridimensionnel. Comme nous pouvons le constater, le paramètre p affecte

considérablement le résultat de la régularisation. Nous observons la différence entre la figure 3.1-b avec $p = 2$ et la figure 3.1-c avec $p = 0.5$. Plus de simplifications du graphe sont obtenues pour $p < 1$. Dans ce cas, La forme du nuage est plus compacte, plus claire et le processus de classification est amélioré quand p diminue.

Deuxième partie

Applications

Chapitre 4

Un cadre unifié pour la reconnaissance des comportements humains

Sommaire

Le contenu de ce chapitre a été publié dans Elsevier-Neurocomputing (Impact Factor 1.580) : http ://www.sciencedirect.com/science/article/pii/S0925231212003220

Résumé : Dans ce chapitre, nous présentons un cadre unifié pour l'analyse des bases de données vidéo à l'aide des marches aléatoires sur graphe associé à une chaîne de Markov spatio-temporelle. Le cadre proposé constitue une approche efficace pour le clustering, l'organisation des données, la réduction de dimension et de la reconnaissance. Le but de notre travail est de développer une approche visuelle pour la reconnaissance du comportement humain. Notre contribution réside dans trois volets. Tout d'abord, nous utilisons les moments de Zernike spatio-temporels pour coder l'objet d'intérêt présent dans un clip vidéo. Ensuite, nous proposons une nouvelle méthode pour représenter la base de données vidéo comme un graphe non orienté pondéré où chaque sommet est un clip vidéo. Le poids d'une arête entre deux clips vidéo est défini par un noyau gaussien sur leurs moments de Zernike spatio-temporels ainsi que sur leurs voisinages respectifs dans l'espace des caractéristiques. Notre objectif est d'obtenir un espace robuste de faible dimension en analysant les propriétés spectrales du graphe ce qui fournit une transcription efficace des points clés dans une variété euclidienne et permet d'atteindre une précision de classification plus élevée. Enfin, nous décrivons un cadre variationnel pour le débruitage de variétés basé sur le p-laplacien, réduisant ainsi l'impact négatif des valeurs aberrantes, améliorant la classification points-clés et ainsi, augmenter la précision de la reconnaissance. La méthode proposée est testée sur la base des vidéos Weizmann, la base d'actions humaines KTH et sur un corpus de gestes de la main. Les résultats obtenus en utilisant les moments de Zernike spatio-temporels prouvent que la méthode proposée peut effectivement capturer la forme des comportements avec des moments de faible ordre. En outre, notre framework permet de classer les différents comportements et atteint un taux de reconnaissance important.

4.1 Introduction

Dans de nombreuses applications, telles que la vidéo surveillance, la détection des comportements inhabituels ou la traduction en langue des signes, il est important de reconnaître l'activité de l'humain afin d'interpréter son comportement. Ce dernier peut être définie comme une succession temporelle d'actions primitives exécutées par le sujet humain le long du clip vidéo et qui peuvent être composées pour former une activité complexe. La reconnaissance des faits et gestes demeure un problème de recherche stimulant. La plupart des méthodes proposées pour traiter ce problème sont basées sur le calcul des modèles d'apparence, 2D ou 3D, à partir d'une silhouette, où une tâche importante consiste à identifier les différentes parties du corps comme la tête, les mains, les pieds et les articulations. D'autres méthodes cherchent à contrôler et interpréter le comportement humain en utilisant les techniques d'estimation du mouvement et les techniques de description tels que le flux optique.

4.2 État de l'art

Pour reconnaître le comportement humain, Ali et al. [Anjum, 2001] utilisent les silhouettes des personnes pour classer une série continue d'actions par l'extraction des propriétés du squelette à partir de la forme. Les caractéristiques du squelette étoile ont été introduites par Fujiyoshi et al. [Fujiyoshi, 1998] pour extraire la posture 2D à partir d'une silhouette en temps réel. Une nouvelle caractéristique, distance étoile, est définie de telle sorte que les vecteurs de caractéristiques peuvent être traduits en symboles par quantification vectorielle. Sur la base de cette distance, la classification des actions est assurée par modèles de Markov cachés (HMM). Afin de reconnaître les actions, Ahmad et al. [Ahmad, 2006] proposent d'extraire le flot optique et les caractéristiques de la forme corporelle de multiples points de vue. La silhouette d'une personne est représentée par un sous-espace réduit en utilisant ACP. Chaque action est représentée par un ensemble de HMMs discrets multidimensionnels modélisés indépendamment de toute direction d'observation. Lawrence et al. dans [Lawrence, 2003,Lawrence, 2005], ont introduit le modèle GPLVM (the Gaussian Process Latent Variable Model) pour la réduction non linéaire de la dimensionnalité permettant, ainsi, la visualisation des données de haute dimension. Contrairement à l'ACP, cette méthode effectue le mapping des données de l'espace réduit (espace latent) vers l'espace de données. Dans [Henrik, 2008], ce modèle est utilisé pour estimer une posture articulée 3D de l'humain à partir de ses différentes silhouettes. Inspirés par [Lawrence, 2005], Wang et al. dans [Wang, 2008] ont introduit le modèle GPDM (the Gaussian Process Dynamic Model) qui comprend un espace latent de faible dimension muni de sa propre

dynamique, et un mapping de ce dernier vers un espace d'observation. Effectivement, GPLVM et GPDM peuvent réaliser une transformation non linéaire entre l'espace paramétré par les mouvement humaines et l'espace latent en réalisant un mapping inverse. Ces deux modèles permettent la description du mouvement humain dans un espace latent de faible dimension. Urtasun et al. dans [Urtasun, 2006] utilisent les modèles GPLVM et GPDM pour apprendre des modèles a priori et ce, pour le suivi de différents styles de marche 3D de l'être humain. Ils ont obtenu de bons résultats même en cas de grandes occlusions.

Efros et al. dans [Efros, 2003], comparent deux actions en se basant sur des caractéristiques extraites à partir des mesures du flot optique dans l'espace spatio-temporel. Manor et Irani [Zelnik, 2001] proposent des distributions multi-échelles du gradient temporel pour isoler et regrouper les événements dans de longues séquences vidéos.

Pour détecter les événements vidéos, Zhu et al. [Zhu, 2009] proposent un descripteur spatio-temporel composé des attributs de bas niveau de l'image, tel que les gradients de l'image, les flots optiques afin de capturer les caractéristiques des actions en fonction de leurs apparences et de leurs modèles de mouvement dans un cube d'espace-temps. Un ensemble de sacs à mots (bag-of-words) (BoW) est construit à partir de ce descripteur à plusieurs niveaux de résolution pyramido-spatiale.

Dans [Laptev, 2004], Laptev et al. comparent deux actions en égalant leurs points d'intérêt (les points Harris). Blank et al. [Blank, 2005] utilisent une pile de points, où les silhouettes sont extraites et évaluées à l'aide de l'équation de Poisson pour chaque point. Bobick et Davis suggèrent l'utilisation des images de l'énergie de mouvement (MEI) et celles de l'histoire de mouvement (MHI) [Bobick, 2001] pour représenter comment une action est réalisée en utilisant différents niveaux d'intensité en fonction du temps écoulé depuis que la silhouette a été capturée.

Sur un autre plan, les méthodes proposées pour la reconnaissance des gestes de la main sont principalement divisées en deux approches. Celles basées sur les gants de données et qui utilisent des dispositifs de détection pour la numérisation de la main et des doigts, et celles basées sur la vision qui nécessitent, seulement, l'utilisation d'une caméra.

Les premières méthodes dédiées au problème de reconnaissance des gestes de la main consistent à détecter la présence de la couleur des marqueurs sur les doigts afin d'identifier ceux qui sont actifs dans le geste. Une revue des méthodes existantes pour l'interprétation des gestes de la main est présentée dans [Pavlovic, 1997]. Des méthodes récentes basées sur des techniques avancées de vision par ordinateur ne nécessitent pas de marqueurs. Concrètement, et d'un point de vue différent, ces méthodes peuvent

être divisées en deux autres approches principales : d'une part, les approches basées sur les modèles 3D des mains [Wu, 2001], qui nécessitent l'utilisation des modèles géométriques (mesh) et des techniques d'animation pour capturer les articulations des mains et ses mouvements. D'autre part, les approches basées sur l'apparence [Chen, 2007], qui utilisent les caractéristiques de l'image pour modéliser l'aspect visuel du geste. D'autres outils de vision par ordinateur, utilisés pour la reconnaissance de gestes 2D et 3D, comprennent des architectures de transformation spécialisées [Rosales, 2001] et des filtres à particules [Bretzner, 2002].

4.3 Solution proposée

Dans ce chapitre nous développons un nouveau framework constituant une approche visuelle pour la reconnaissance du comportement humain. Le cadre proposé se compose de trois phases qui peuvent être exécutées séparément mais séquentiellement. Tout d'abord, nous utilisons les moments de Zernike spatio-temporels pour décrire l'objet d'intérêt (OI) dans le clip vidéo. Ces descripteurs sont invariants aux changements d'échelle, de rotation et de translation. En outre, ils permettent de récupérer des informations à la fois temporelles et spatiales. L'extraction d'un OI a ses propres problèmes. Par conséquent, nous supposons qu'il a été déjà segmenté et qu'il représente une seule personne. Après cela, nous proposons un modèle de diffusion pour la réduction de la dimensionnalité qui fournit un cadre solide et efficace pour générer de nouvelles coordonnées pour les vidéos et les classer dans un espace euclidien.

Une fois le cadre de diffusion défini, nous allons utilisé le cadre variationnel décrit dans le chapitre 3 afin d'améliorer la classification des OI, réduisant ainsi l'impact négatif des valeurs aberrantes sur le processus de catégorisation. Les principales nouveautés de notre travail sont : 1) la prise en compte de la dimension temps dans les moments de Zenike et cela pour la description d'une entité temporelle et 2) l'approche proposée est plus robuste (moins sensible aux particularités des données ou au bruit) permettant une précision plus élevé de classification.

Le framework proposé sera validé par deux applications : 1) catégorisation des actions humaines, et 2) une approche visuelle pour l'identification des propriétés d'un objet manipulé par les mains. L'idée de base de notre travail est de considérer toutes les vidéos comme un graphe pondéré, où ses sommets (clips vidéo) sont représentés par des volumes 3D (les descripteurs de Zernike spatio-temporels), et ses arêtes représentent la similarité entre les sommets connectés. Le poids d'une arête entre deux clips vidéo est défini par un noyau gaussien sur leurs moments de Zernike spatio-temporels ainsi que sur leurs voisinages respectifs, et ce dans l'espace des caractéristiques. Les points

saillants de notre framework basé sur l'analyse des propriétés spectrales du graphe, sont les suivants :

1. Le regroupement spectral ramène le problème de classification à un problème de partitionnement de graphe où la décomposition spectrale de la matrice associée aux marches aléatoires sur ce graphe permet de définir une variété euclidienne de faible dimension. L'algorithme de regroupement proposé permet de capturer et d'exploiter les similarités entre les clips vidéos dans un graphe k-pp et cela d'une façon dynamique et sans apprentissage.
2. La robustesse de notre modèle est améliorée en régularisant le graphe de voisinage dans l'espace réduit, ce qui permet de débruiter et de simplifier les données. A ce stade, nous pouvons sélectionner un ensemble "fiable" de voisins pour chaque sommet dans son voisinage non local. Ainsi, le clustering sera effectué dans une variété euclidienne régularisée.

Il convient de mentionner qu'à chaque étape, nous allons utiliser un graphe différent (c-à-d : un graphe dans l'espace des caractéristiques pour permettre le regroupement spectral, et un graphe dans l'espace réduit pour la régularisation de la variété inférée). Des tests ont été menés sur trois ensembles de données vidéo. Les résultats d'évaluation des performances obtenus montrent que le framework que nous avons proposé permet de classer efficacement les vidéos et atteindre un très haut taux de reconnaissance.

Dans la section suivante nous allons présenter et commenter les résultat de notre proposition .

4.4 validation expérimentale

Pour évaluer la fiabilité de notre approche, nous considérons deux applications : 1) la catégorisation des actions humaines, et 2) une approche visuelle pour l'identification des propriétés des objets manipulés par les mains. Nous menons nos expériences sur trois bases de données vidéos. Deux d'entre elles sont consacrées à la catégorisation des actions humaines, à savoir : la base KTH [Laptev, 2004] et la base Weizmann [Blank, 2005]. La troisième base, la base EPHE, contient des vidéos clips des gestes de la main et est la propriété de "Ecole Pratique des Hautes Etude, Sorbonne". Les bases KTH et Weizmann sont largement étudiées sous divers aspects et plusieurs résultats sont disponibles. Par conséquent, elles offrent un repère intéressant pour évaluer nos résultats. Par ailleurs, l'ensemble de données des gestes de la main permet de tester notre cadre sous un autre aspect, et donc de consolider les résultats obtenus. Rappelons que notre

framework est composé de trois phases où, d'abord, un volume binaire représentant l'objet d'intérêt (OI) est caractérisé par l'utilisation des moments de Zernike spatio-temporels. Ensuite, une représentation euclidienne de dimension réduite est calculée par regroupement spectral et ce en exploitant les propriétés spectrales du graphe de similarité entre les moments de Zernike. Enfin, pour améliorer la précision de la classification, une régularisation de ce graphe dans l'espace réduit est effectuée.

4.4.1 Catégorisation des actions humaines

Les silhouettes des OI sont fournies avec les deux bases de données KTH et Weizmann. Ainsi, nous pouvons les décrire directement en utilisant les descripteurs 3D.

4.4.1.1 La base de vidéos Weizmann

Cette base de données contient 90 clips vidéo réalisés par des personnes différentes. Chaque clip vidéo contient une seule personne effectuant une action. Il y a dix catégories d'actions représentées dans cette base, à savoir : *walk, run, skip, jack, jump, jump in place, side, wave with one hand, wave with two hands, bend*. Ces catégories sont désignées respectivement ci-après par (a1, a2, a3, a4, a5, a6, a7, a8, a9, a10).

La figure 4.1 montre quelques extraits de cet ensemble de données.

Le choix de l'ordre approprié parmi les moments de Zernike spatio-temporels est crucial pour décrire le comportement du sujet et par conséquent, acquérir plus ou moins de détails sur le volume binaire vidéo. Une approche commune consiste à choisir l'ordre des moments qui permet une meilleure reconstruction de l'objet d'intérêt. Dans [Boyce, 1983], une image en niveaux de gris a été reconstruite en utilisant les moments de Zernike d'ordre croissant. Il a été démontré que l'utilisation des moments de Zernike d'ordre 6 conduit à une reconstruction avec une erreur de 10%. Cette erreur était de 6% en utilisant les moments de Zernike d'ordre 20. Toutefois, ce résultat peut varier en fonction du cas étudié, et l'ordre optimal dépend de la nature des objets à reconstruire et ne peut donc pas être généralisé. Par ailleurs, en terme de sensibilité au bruit additif aléatoire, il a été montré dans [Teh, 1988] que, en présence du bruit, l'image est reconstruite à l'aide de moments allant jusqu'à un certain ordre optimal. La reconstruction de l'image à l'aide de moments d'ordre supérieur à l'ordre optimal dégrade sa qualité, car les moments d'ordre supérieur sont plus vulnérables au bruit blanc.

Dans cette application, nous nous intéressons principalement à la reconnaissance des activités humaines et par conséquent, à la performance globale de l'approche proposée. Ainsi, nous avons testé différents ordres pour sélectionner un ordre approprié des

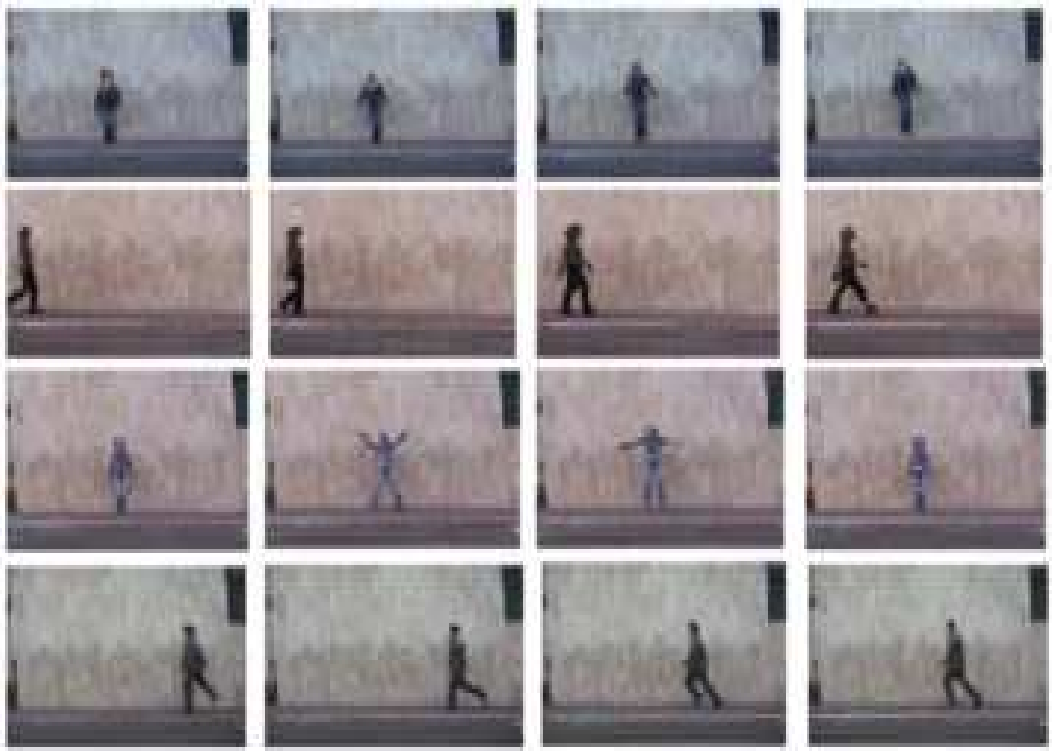

FIGURE 4.1 – Exemple d'actions de la base Weizmann

moments qui permet un meilleur taux de reconnaissance. Avant de procéder à la phase de régularisation, la Table 4.1 montre les variations des taux de reconnaissance avec différents ordres de moments. Comme nous pouvons le voir, le taux de reconnaissance obtenu en utilisant les moments d'ordre 7 n'a pratiquement pas de différence avec celui obtenu en utilisant les moments d'ordre 6.

	Taux de reconnaissance des actions (%)										
ordre des moments	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	**Précision moyenne**
Ordre7	93.1	95	87.2	96	79	93	89	91.3	93.2	94	**91,08**
Ordre6	92.5	94.8	86.4	95.9	78.2	93	88.8	90.6	92.9	93.7	**90.68**
Ordre5	91.1	90.7	85.9	92	77.3	84.8	84.5	85.2	87.1	91	**86.96**
Ordre4	88.7	86	84.7	85	75	76.3	79.3	78	80.7	88.3	**82.2**
Ordre3	85	80	83	75	70	65	73	68	71	81	**75.1**

TABLE 4.1 – Comparaison des taux de reconnaissance des actions avec différents moments

En utilisant les moments d'ordre 7, et après la phase de régularisation dans l'espace réduit, nous obtenons la matrice de confusion pour les 10 actions de cet ensemble de données, présentée dans la Table 4.2.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	100	0	0	0	0	0	0	0	0	0
a2	2	98	0	0	0	0	0	0	0	0
a3	0	3	97	0	0	0	0	0	0	9
a4	0	0	0	100	0	0	0	0	0	0
a5	10	0	0	0	83	7	0	0	0	0
a6	0	0	2	0	0	98	0	0	0	0
a7	0	0	0	0	0	0	96.3	1.3	2.4	0
a8	0	0	0	0	2	0	0	96	0	2
a9	0	0	0	0	0	0	0	4	96	0
a10	0	0	0	0	0	0	0	1	0	99

TABLE 4.2 – La matrice de confusion en utilisant les moments Zernike spatio-temporels de l'ordre 7.

Il est assez clair que la régularisation discrète du graphe dans l'espace réduit améliore de façon significative les taux de reconnaissance des actions. Dans l'ensemble, la précision moyenne est de **96,33%**, alors que pour le même ordre, sans régularisation, la précision moyenne était de **91,08 %**. Les résultats obtenus à partir de cette expérience sont comparés à ceux rapportés dans d'autres travaux ([Zelnik, 2001, Xinghua, 2009, Vezzani, 2010, Kellokumpu, 2011, Dhillon, 2009]) (voir la Table 4.3).

De cette comparaison, il s'avère que notre méthode est compétitive avec les méthodes de l'état de l'art, et permet d'obtenir des résultats encourageants par rapport à ceux publiés antérieurement.

4.4.1.2 La base KTH

Nous avons testé notre approche sur une deuxième base de vidéos (the KTH human motion dataset). Elle contient six types d'actions humaines (walking, jogging, running, boxing, hand waving, and hand clapping). Chacune de ces actions est effectuée par 25 personnes dans quatre situations différentes (à l'extérieur, à l'extérieur avec des variations d'échelle, en plein air avec des vêtements différents, et à l'intérieur). La figure 4.2 montre quelques extraits de cet base de données.

Après avoir caractérisé les OIs par les moments de Zernike spatio-temporels, une projection dans une variété euclidienne est réalisée et des coordonnées de faible dimension sont déduites. La projection des clips vidéo dans cet espace réduit par l'utilisation de ces nouvelles coordonnées est illustrée dans la figure 4.3.

	Taux de reconnaissance des actions (%)										
Méthode	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	**Précision moyenne**
Notre approche	100	98	97	100	83	98	96.3	96	96	99	**96.33**
Kellokumpu et al. [Kellokumpu, 2011]	100	100	100	100	89	100	100	100	100	100	**98.9**
Xinghua et al. [Xinghua, 2009]	100	90	90	100	100	100	100	100	100	100	**97.8**
Dhillon et al. [Dhillon, 2009]	91	93	69	94	92	92	90	91	92	95	**89.9**
Vezzani et al. [Vezzani, 2010]	100	99	68	87	81	95	57	100	86	94	**86.7**
Zelnik et al. [Zelnik, 2001]	82.4	34.7	43.5	95.5	29.2	84.9	50.8	29.6	51.9	86.6	**58.91**

TABLE 4.3 – Comparaison des performances sur la base Weizmann

FIGURE 4.2 – Exemple d'actions de la base KTH

Bien que l'ensemble des données vidéo est clairement classés, quelques ambiguïtés subsistent encore. Ces ambiguïtés sont mieux séparées dans la figure 4.4, qui projette les mêmes clips vidéo dans l'espace réduit régularisé.

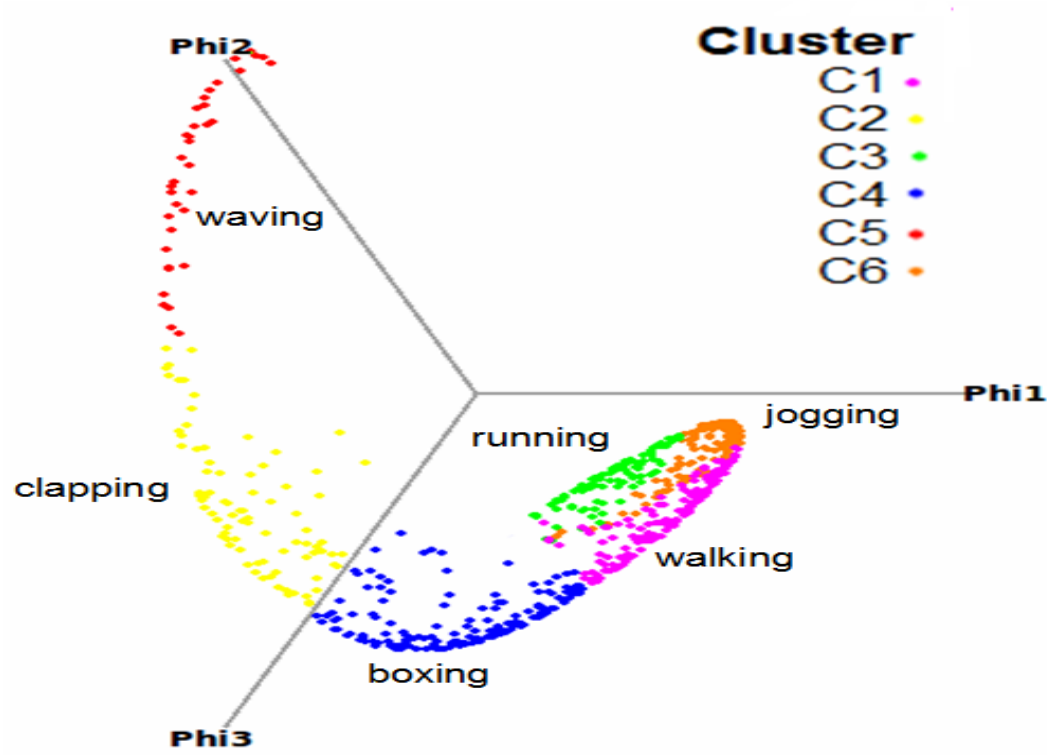

FIGURE 4.3 – Les actions de la base KTH projetées dans l'espace réduit

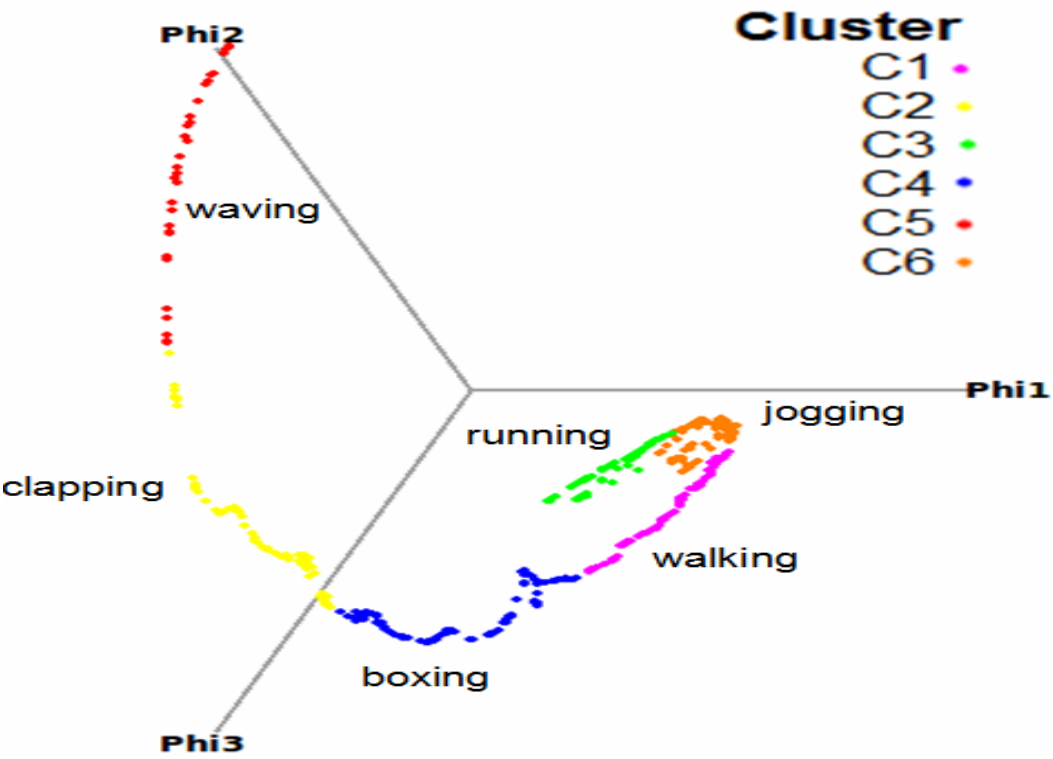

FIGURE 4.4 – Les actions de la base KTH projetées dans l'espace réduit régularisé

	bx	jg	rg	wg	hd-cl	hd-wa
boxing	95.6/**98**	0	0	3/2	1.4/0	0
jogging	1.6/0	84.2/**89**	6.2/6	8/5	0	0
running	0	3/1	95.4/**98**	1.6/1	0	0
walking	1/0	2/0	0	97/**100**	0	0
hand-clapping	6/5	0	0	0	89.5/**92**	4.5/3
hand-waving	5.6/4	0	0	0	5/2	89.4/**94**

TABLE 4.4 – La matrice de Confusion sans et avec régularisation.

La table 4.4, illustre la matrice de confusion relative à cette base de données. Nous pouvons constater que les résultats obtenus avec régularisation sont considérablement meilleurs. Dans l'ensemble, la précision moyenne est de **95.17%**.

Pour confirmer la fiabilité du framework que nous avons proposé, les résultats obtenus par cette expérience sont également comparés à ceux obtenus avec d'autres méthodes de l'état de l'art. ([Xinghua, 2009, Kellokumpu, 2011, Dhillon, 2009, Costantini, 2011, Ballan, 2009] (voir la figure 4.5 et la Table 4.5). Comme nous pouvons le voir, les résultats sont comparés favorablement et notre contribution est contrastée.

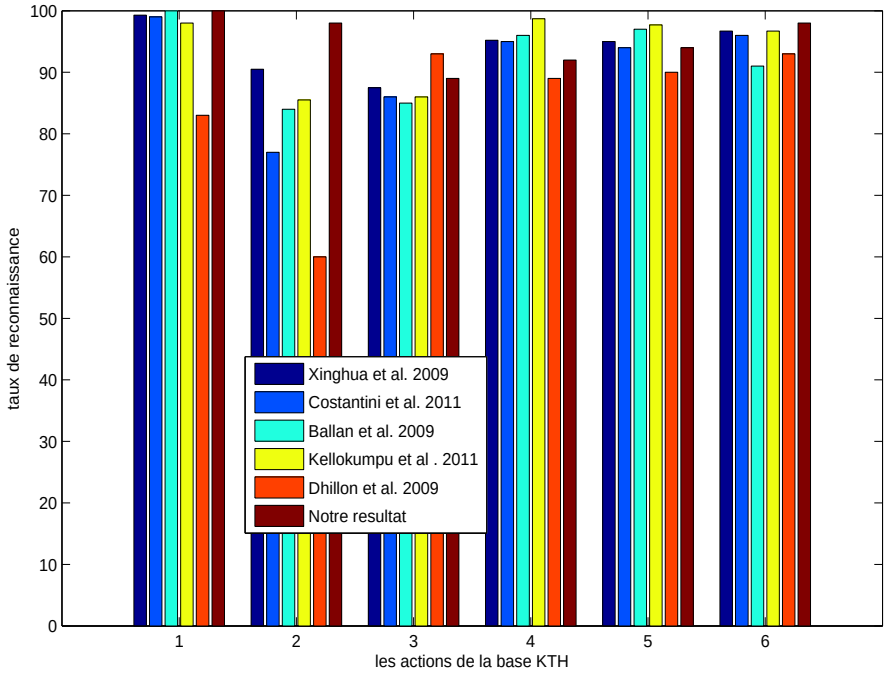

FIGURE 4.5 – La reconnaissance des actions en utilisant notre approche comparée avec d'autres méthodes de l'état de l'art (la base KTH)

Méthodes	Taux (%)
Notre approche	95.17
Xinghua et al. [Xinghua, 2009]	94.0
Costantini et al. [Costantini, 2011]	91.17
Ballan et al. [Ballan, 2009]	92.17
Kellokumpu et al. [Kellokumpu, 2011]	93.77
Dhillon et al. [Dhillon, 2009]	84.67

TABLE 4.5 – Comparaison des performances sur la base KTH

4.4.2 Une approche visuelle pour l'identification des propriétés des objets

De la même manière, nous avons également testé notre approche pour la reconnaissance des gestes de la main. Le problème clé, dans ce contexte, est de rendre les gestes de la main compréhensibles par les ordinateurs afin de reconnaître certaines propriétés des objets et cela en utilisant seulement une caméra. Dans cette application, nous cherchons à reconnaître la texture et la consistance d'un objet à travers l'analyse vidéo des actions de la main. Deux propriétés de l'objet sont étudiées : la texture qui peut être soit lisse ou granuleuse et la consistance qui peut être soit dure ou molle. Nous nous sommes inspirés de la définition fondamentale de Lederman et Klatzky [Lederman, 1987], et nous avons étudié quatre procédures exploratoires :

1. Mouvement latéral sur objet lisse (Lateral Motion for Smooth Object *LMSO*).
2. Mouvement latéral sur objet granulé (Lateral Motion for Granular Object *LMGO*).
3. Pression sur objet mou (Pressure for Soft Object *PSO*).
4. Pression sur objet dur (Pressure for Hard Object *PHO*).

Nous avons effectué des expériences sur le corpus EPHE composé de 21 clips vidéo de gestes représentant la manipulation de différents objets par, à la fois, la main gauche et la main droite. La figure 4.6 montre les objets manipulés et quelques images extraites des clips vidéo.

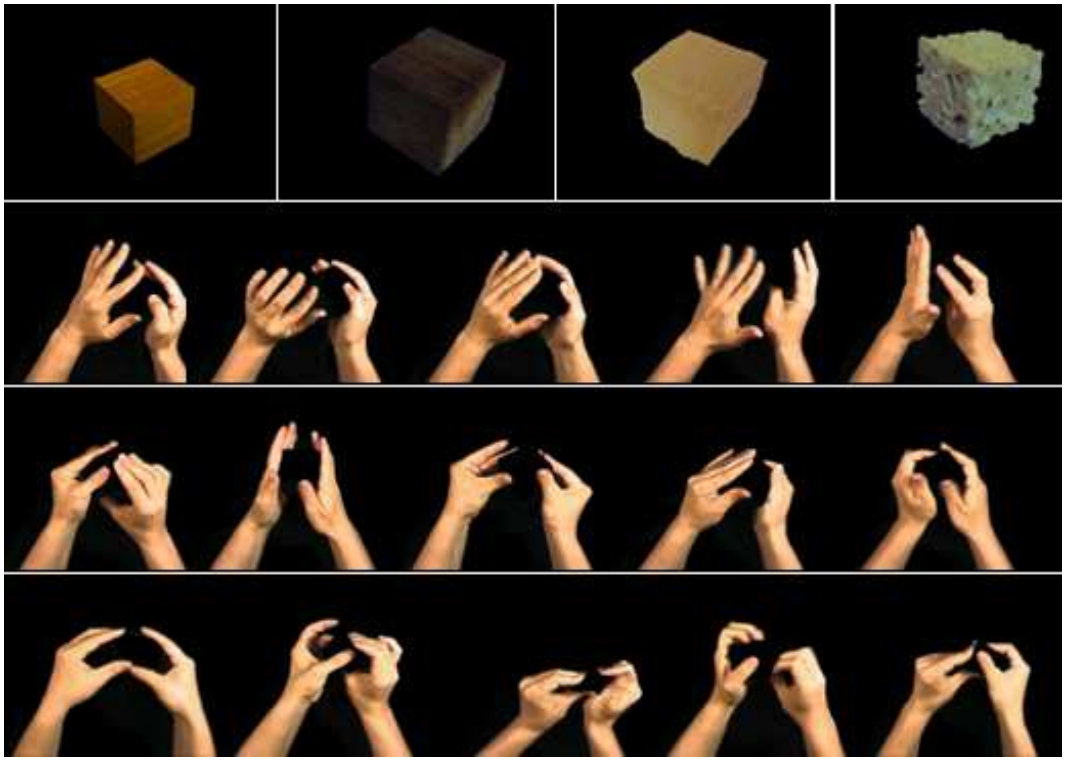

FIGURE 4.6 – Les objets manipulés et quelques images extraites des clips vidéo

FIGURE 4.7 – Les mains gauches et les mains droites ordonnées par le vecteur de Fiedler

Pour vérifier l'efficacité des moments de Zernike spatio-temporels dans notre approche, nous avons appliqué le même processus séparément sur chaque main (main gauche / main droite). La figure 4.7 montre le résultat de projection des mains séparées en utilisant le vecteur de Fiedler, ce qui permet de classer, respectivement, les mains gauche et droite. Pour la classification des gestes, nous avons opté pour la séparation des deux mains. Chaque main est représentée par ses moments 3D. La distance entre deux gestes est calculée par : $\frac{\|f(v)^{droite} - f(u)^{droite}\| + \|f(v)^{gauche} - f(u)^{gauche}\|}{2}$.

Le résultat de la reconnaissance de la main gauche et de la main droite dans ces vidéos est de 100%. Les taux de reconnaissance des gestes dans les clips vidéo sont résumés dans la Table 4.6. Une fois de plus, la régularisation discrète du graphe dans l'espace réduit permet d'obtenir de meilleurs résultats. Cela consolide les résultats obtenus dans la première application et par conséquent, confirme l'efficacité de notre approche.

	taux de reconnaissance (%)	
Geste reconnu	sans régularisation	avec régularisation
LMSO	90	93.6
LMGO	85.7	89.2
PSO	71.4	77.4
PHO	80	85.9

TABLE 4.6 – Reconnaissance des gestes avec et sans régularisation

4.5 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle méthode pour la reconnaissance du comportement humain qui repose sur une extension des moments de Zernike au domaine spatio-temporelle et cela pour caractériser les actions de l'être humain dans les clips vidéo. En effet, les moments de Zernike spatio-temporels présentent des propriétés intéressantes pour décrire l'information structurelles et temporelles de la séquence vidéo. Notre approche se compose de trois étapes principales. La première utilise les moments de Zernike spatio-temporels pour décrire la silhouette et la dynamique de l'objet d'intérêt (OI) dans le clip vidéo. Ensuite, dans la deuxième étape, nous construisons un graphe de similarité visuelle en utilisant les caractéristiques 3D. L'idée de base de cette étape est de considérer toutes les vidéos comme un graphe pondéré dont les sommets (clips vidéo) sont représentés par les volumes 3D (descripteurs de Zernike spatio-temporels) et les arêtes représentent la similarité entre les sommets connectés. Enfin, nous utilisons une approche variationnelle pour le débruitage de variétés, ce qui nous permet d'exploiter la géométrie et la distribution des données.

Nous avons validé notre approche à travers deux applications : 1) la catégorisation des actions humaines, et 2) une approche visuelle pour l'identification des objets manipulés par les mains. À la lumière des résultats obtenus pour les deux applications, il s'avère que notre framework permet d'obtenir des taux de reconnaissance significatifs par rapport aux autres méthodes de l'état de l'art et que les moments de Zernike spatio-temporels peuvent effectivement capturer la forme des comportements avec des moments d'ordre faible, ce qui confirme l'efficacité de l'approche proposée.

Chapitre 5

Apprentissage faiblement supervisé à partir des points SIFT

Sommaire

Le contenu de ce chapitre va être publié dans SPIE-Journal of Electronic Imaging (5-Year Impact Factor : 0.694)

Résumé : Dans ce chapitre nous proposons une approche unifiée de diffusion de labels dans un espace de grande dimension en utilisant un ensemble restreint de points SIFT. Notre contribution comporte quatre volets. Premièrement, nous identifions automatiquement les points saillants de l'espace à explorer en utilisant les points clés de l'image. Ensuite nous construisons un graphe sur ses points et à partir du spectre de la matrice relative aux marches aléatoires sur ce graphe, nous proposons une projection des point clés dans un espace euclidien de faible dimension. Après cela, nous utiliserons le cadre variationnel, décrit dans le chapitre 3, pour déduire une structure compacte du graphe dans l'espace réduit afin d'améliorer la classification de ces points-clé. Enfin, nous décrivons notre algorithme de diffusion multi-label sur graphe moyennant le même cadre variationnel. Une analyse théorique de cet algorithme est développée ainsi que des liens avec d'autres méthodes de la littérature. Les résultats de la segmentation montrent que notre approche permet de propager efficacement les connaissances initiales et d'obtenir de bons résultats.

5.1 Introduction

Les algorithmes d'apprentissage utilisant les graphes se sont répandus ces dernières années. Dans ces algorithmes, des informations initiales (des labels) sont demandées à l'utilisateur pour mettre en évidence les points critiques de l'espace de données à explorer. La diffusion de ces labels sur le graphe permet de catégoriser la totalité de l'ensemble des données. En conséquence, les résultats de cette classification dépendent de l'initialisation. L'évitement des interactions manuelles avec l'utilisateur et la propagation des connaissances initiales sera d'un grand intérêt, dans le domaine de la vision par ordinateur et dans ses domaines connexes, où la segmentation d'objets dans les images est un sujet de recherche ouvert.

5.2 État de l'art

Dans la littérature, nous trouvons différentes approches de segmentation d'images utilisant les graphes. Parmi lesquelles, une grande famille d'algorithmes sont basés sur une minimisation de fonctionnelles d'énergie où, généralement, l'image est modélisée par un graphe et l'objet à extraire est celui pour lequel une certaine forme de fonctionnelle d'énergie doit être minimisée [Boykov, 2001, Kolmogorov, 2004]. Deux tendances de segmentation se distinguent pour extraire les objets d'intérêt :

- le processus de segmentation se fait automatiquement ou,
- il est guidé par des informations fournies par l'utilisateur.

Pour effectuer une segmentation automatique, Campbell et al. [Campbell, 2010], utilisent une méthode itérative basée sur les graph-cut ainsi que sur le point de fixation de la caméra sur l'objet d'intérêt pour en déduire un modèle de couleur et déterminer un contour englobant. Dans une approche similaire, Kim et Hong [Kim, 2008] minimisent une fonctionnelle d'énergie exprimée par un modèle de contour actif où l'image est divisée en deux régions à l'aide d'un graph-cut.

Dans l'approche interactive, l'utilisateur fournit des germes initiaux pour différentier entre l'objet d'intérêt et le fond et selon l'emplacement des germes, deux types d'initialisation sont étudiées. Soit elle est basée sur une estimation de l'objet et de l'arrière-plan [Boykov, 2001, Rother, 2004], soit elle commence par un contour initial [Xu, 2007], qui peut entourer l'objet ou une partie [Blake, 2004], ou bien elle délimite la frontière entre l'objet et le fond [Wang, 2007].

Cependant, cette approche n'est pas efficace si le processus de segmentation est intégré dans un cadre automatisé. En outre, Le résultat est très dépendant de l'initialisa-

tion. En effet, il est possible que parmi les germes introduits par l'utilisateur pour désigner des parties de l'objet (resp. arrière-plan). Certains d'entre eux sont de la même classe de l'arrière-plan (resp. objet)(ex. l'image à segmenter est complexe). Aussi, il se peut que les germes ne soient pas bien répartis sur l'image et par conséquent, ne soient pas assez représentatifs.

5.3 Solution proposée

Nous proposons dans ce chapitre une alternative à l'interaction manuelle en utilisant les points caractéristiques comme vérité terrain. Dans un premier temps, ils sont régularisés et classés avant d'être diffusés comme labels sur l'image initiale pour en extraire les objets d'intérêt. Une variété de descripteurs ont été proposés comme points caractéristiques parmi lesquels les points SIFT (cf. 2.2.1.2).

Cependant, chaque point SIFT est décrit par un nombre considérable d'attributs et donc réparti dans un espace de grande dimension. Pour surmonter cette contrainte, nous construisons un graphe de similarité à partir de ces points. La décomposition spectrale de la matrice qui lui est associée permet d'identifier les axes qui sont porteurs des informations les plus pertinentes. Une distance appropriée est alors définie entre ces axes, telle que celle proposée dans [Lafon, 2006], pour capturer la similarité et la proximité entre les points clés. Ensuite, pour faciliter le processus de catégorisation, une régularisation discrète par p-Laplacien est faite sur le graphe construit à partir des points-clé SIFT exprimés dans leurs nouvelles coordonnées de l'espace réduit. Enfin, nous proposons un algorithme de diffusion multi-label sur le graphe de l'image initiale.

– Les grandes lignes de la solution proposée

L'idée de base de notre approche est de montrer comment est-il possible d'apprendre à partir d'un ensemble restreint de données et de propager les connaissances acquises dans une base de données de grande dimension. Pour évaluer l'efficacité de notre approche, nous étudions dans cette application le cas de la segmentation d'images. Le cadre proposé fonctionne en deux phases. Dans la première phase, des germes sont identifiés et classés automatiquement et dans la deuxième phase, une diffusion de ces germes sur le graphe permet de mettre en évidence l'objet d'intérêt. Les étapes suivantes décrivent notre approche :

– Un ensembe de points SIFT est extrait de l'image et utilisé pour construire un

graphe de similarité visuelle. Une analyse spectrale de ce graphe est réalisée afin de définir un espace euclidien réduit [Lafon, 2006]. Pour accélérer la décomposition spectrale de la matrice associée à ce graphe, nous utilisons la méthode décrite dans 2.3.4. Pour améliorer le processus de catégorisation, la régularisation discrète est effectuée sur le graphe construit sur les points SIFT exprimés dans leurs nouvelles coordonnées. Ainsi, la segmentation est effectuée dans un espace euclidien réduit et régularisé.

– A ce stade, un nouveau graphe est construit sur l'image. Il contient des sommets étiquetés et des sommets non étiquetés. En utilisant notre algorithme de propagation multi-label, une fonctionnelle d'énergie est formulée et minimisée afin d'extraire l'objet d'intérêt.

Il est à noter qu'à chaque étape, nous allons utiliser un graphe différent :

1. Un graphe sur les descripteurs SIFT (espace des caractéristiques) afin de déduire de nouvelles coordonnées de faible dimension dans un espace euclidien.
2. Un graphe dans l'espace réduit. La régularisation de ce dernier graphe améliore la précision de la classification des points SIFT.
3. Un graphe final est construit sur toute l'image (données d'entrées initiales), une fois que les points SIFT sont étiquetés (classés). Ce dernier graphe contient des sommets étiquetés et des sommets non étiquetés. Il sera utilisé pour propager les étiquettes des points SIFT dans leurs voisinages.

Une illustration graphique de notre cadre est présentée dans la figure 5.1. Dans la section suivante nous détaillons notre algorithme de diffusion multi-label ainsi que des liens de cet algorithme avec d'autres méthodes.

5.4 Algorithme de diffusion multi-label

Soit V l'ensemble des pixels de l'image, $V_L = \{v_k\}_{k=1}^{m}$ l'ensemble des points labellisés (points SIFT classés) et $V_U = \{v_u\}_{u=m+1}^{N}$, l'ensemble de ceux qui ne le sont pas encore. Nous étendons la fonction $f(v)$ définie sur le sommet v (voir § 2.2), pour intégrer la valeur de son label $f_0 = l \in \mathscr{L} = \{1,2,...,c\}$. f sera donc représentée par le tuple $\{f_0, f_1, ..., f_q\}$. Pour rendre la similarité entre les sommets du graphe insensible à h_i (cf eq.2.20), nous normalisons chaque $w(u,v)$ comme suit : $w(u,v) = \dfrac{w(u,v)}{\max\limits_{v \sim u} w(u,v)}$.

La propagation de labels peut être formulée comme une minimisation de la fonction d'énergie exprimée par l'équation 3.17. Généralement, on utilise $p = 2$, alors notre stratégie de propagation de labels peut être formulée comme un processus itératif où, à

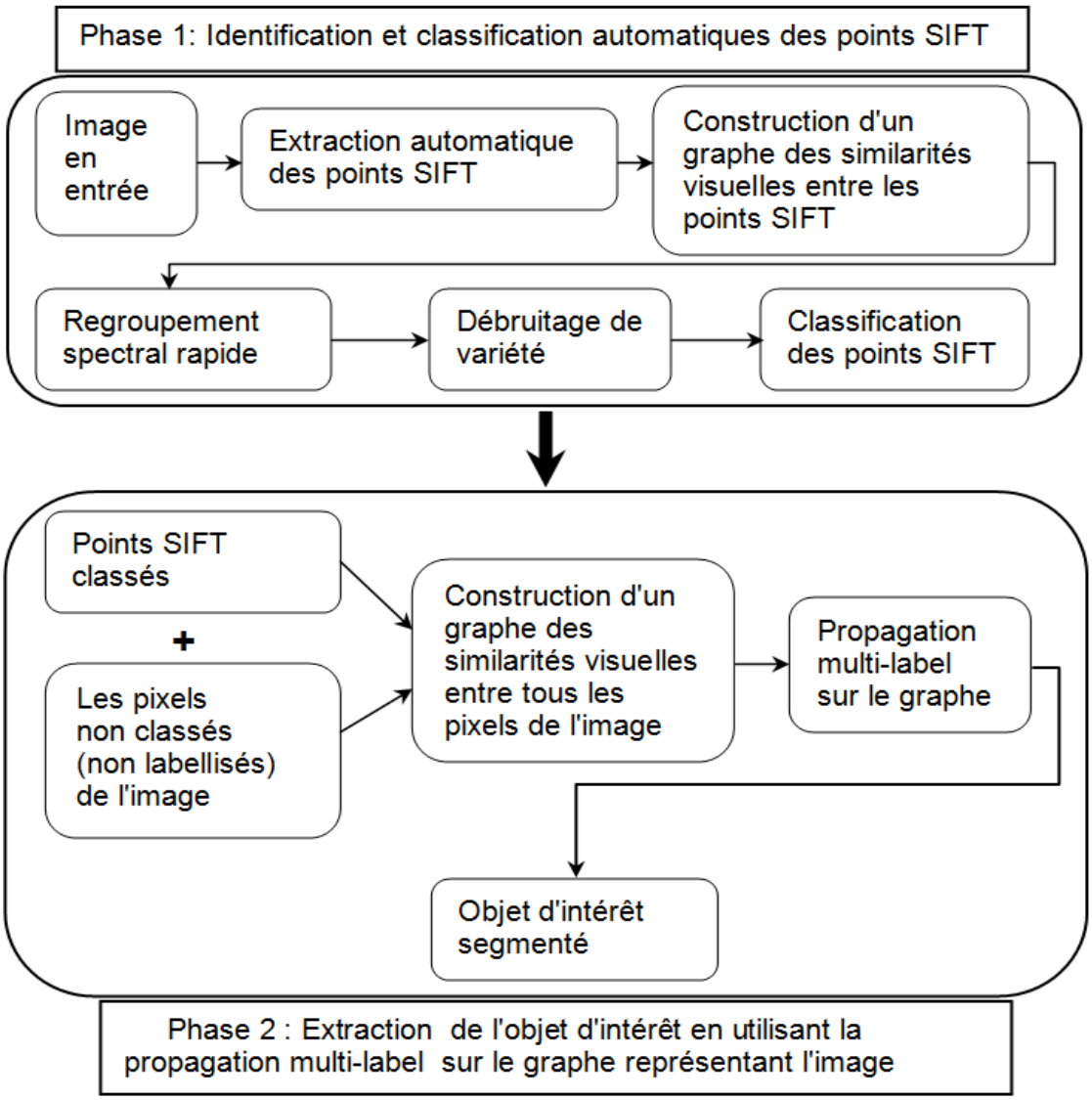

FIGURE 5.1 – Notre cadre de diffusion

chaque itération, seules les étiquettes des sommets non labellisés seront mises à jour. Pour un pixel non labellisé v, son label à l'itération t sera calculé par :

$$\begin{cases} f = (f_0, f_1, \cdots, f_K), \quad f^0 = f_0 \in \mathscr{L} \\ f_v^{t+1} = \dfrac{1}{\lambda + \sum\limits_{u \sim v} w(u,v)} (\lambda f_v^0 + \sum\limits_{u \sim v} w(u,v) f_u^t) \end{cases} \tag{5.1}$$

Si on pose $\lambda = 0$, et $p(u,v) = \dfrac{w(u,v)}{\sum_{v \sim u} w(u,v)}$, on peut définir le processus de propagation à partir d'un sommet u vers un autre sommet v par : $f_v^{t+1} = \dfrac{1}{\sum\limits_{u \sim v} w(u,v)} \sum\limits_{u \sim v} w(u,v) f_u^t =$

$\sum_{u\sim v} p(u,v) f_u^t$. Alors, l'équation (5.1) peut être réécrite comme suit :

$$\begin{cases} f^0 = f_0 \in \mathscr{L} \\ f^{t+1}(v) = \sum_{u\sim v} p(u,v) f^t(u) \quad \forall v \in V \end{cases} \tag{5.2}$$

Ensuite, la procédure de propagation multi-label sur le graphe G peut être vue comme une classification spécifique $\mathscr{C}$ sur V. Elle est considérée comme une fonction qui attribue des labels pour chaque sommet v :

$$f_0(v) = \arg\max_{l\leq c} \mathscr{C}_{vl}. \tag{5.3}$$

Initialement, soit $\mathscr{C}^0_{vl} = 1$ si v est labellisé comme l, et $\mathscr{C}^0_{vl} = 0$ sinon. Pour un sommet non labellisé v, $\mathscr{C}^0_{vl} = 0$. Par conséquent, (5.3) peut être reformulée par :

$$f_0(v) = \arg\max_{f_0(u)\leq c} (f_0(v) = f_0(u)). \tag{5.4}$$

L'idée de base de notre méthode de propagation de labels, est de considérer un algorithme itératif où chaque nœud absorbe une partie des informations sur les labels de son voisinage et met à jour son propre label. Cette procédure sera répétée jusqu'à ce que tous les nœuds du graphe soient labellisés et ne changent plus.

– Liens avec d'autres méthodes

Les algorithmes de segmentation utilisant les graphes ont eu beaucoup de succès ces dernières années. Les variantes modernes sont principalement construites à partir d'un petit ensemble d'algorithmes de base : des coupes de graphes, marche aléatoire, et les algorithmes du plus court chemin. Récemment, ces trois algorithmes ont été placés dans un cadre commun qui leur permet d'être considérés comme un cas particulier d'un algorithme de segmentation semi-supervisée général avec différents choix des paramètres p et q [Sinop, 2007] :

$$\sum_{(u,v)\in E} (w(u,v)^p |f_u - f_v|)^q \tag{5.5}$$

où $w(u,v)$ est une fonction qui mesure les interactions entre les sommets du graphe et $|f_u - f_v|$ mesure la distance entre eux. Ainsi, notre méthode (5.2) peut facilement être déduite de ce cadre si nous posons $p = q = 1$, et $p(u,v) f^t(v) = w(u,v)|f_u - f_v|$. De plus, une connexion entre 5.2 et la minimisation par les modeles MRF (Markov Random field) peut aussi être établie. Rappelons qu'un MRF est souvent décrit par un ensemble

de sommets V régi par une relation donnée. Sur chaque sommet v, il y a une variable aléatoire $f(v)$ qui prend des valeurs d'un ensemble fini ($f(v) \in \mathscr{L} = \{1, 2, ..., c\}$). Le but est de trouver f^* qui satisfait :

$$f^* = \text{argmin}\{\sum_v \phi(f(v)) + \sum_{u \sim v} \phi_{uv}(f(u), f(v))\} \tag{5.6}$$

$\phi(f(v))$ est une fonction sur la variable $f(v)$ qui peut être vue comme une énergie de vraisemblance, où :

$$\phi(f(v)) = \begin{cases} \infty & \text{if } f^{t+1}(v) \neq f^t(v) \\ & \text{(le label de } v \text{ va être be changé)} \\ 0 & \text{autrement} \end{cases} \tag{5.7}$$

et $\phi_{uv}(f(u), f(v))$ mesure l'échange d'informations entre le sommet labellisé u et le sommet non labellisé v. A son tour, elle peut être définie, en se référant à (5.2) et (5.5), par : $\phi_{uv}(f(u), f(v)) = p(u,v)|f(u) - f(v)|$. $p(u,v)$ est la probabilité correspondante à la marche aléatoire de u vers v. Comme v devrait être labellisé (donc, $f(v) = \infty$), alors la minimisation de l'équation (5.6) est équivalente à la résolution du problème d'optimisation suivant :

$$\min_{f(u) \in \mathscr{L}} \left\{ \sum_{u \sim v} p(u,v)|f(u) - f(v)| \right\} \tag{5.8}$$

Ainsi, nous avons montré que notre méthode peut être dérivée du cadre de minimisation d'énergie du MRF.

5.5 Expérimentations

Nous avons effectué nos tests sur une collection d'images de la base d'images Berkeley [Martin, 2001]. D'abord, des points SIFT sont situés sur chaque image, puis un graphe de similarité visuelle est construit à partir de ces points-clés en utilisant la mesure de similarité w_{uv} (cf. eq. 2.20). $F(.)$ utilisé, est l'histogramme du patch entourant le point considéré exprimé dans l'espace LAB. Ensuite, la décomposition spectrale de la matrice de transition, déduite du dernier graphe, permet de définir un nouvel espace réduit où chaque point SIFT est exprimé par de nouvelles coordonnées calculées à partir du patch qui l'entoure.

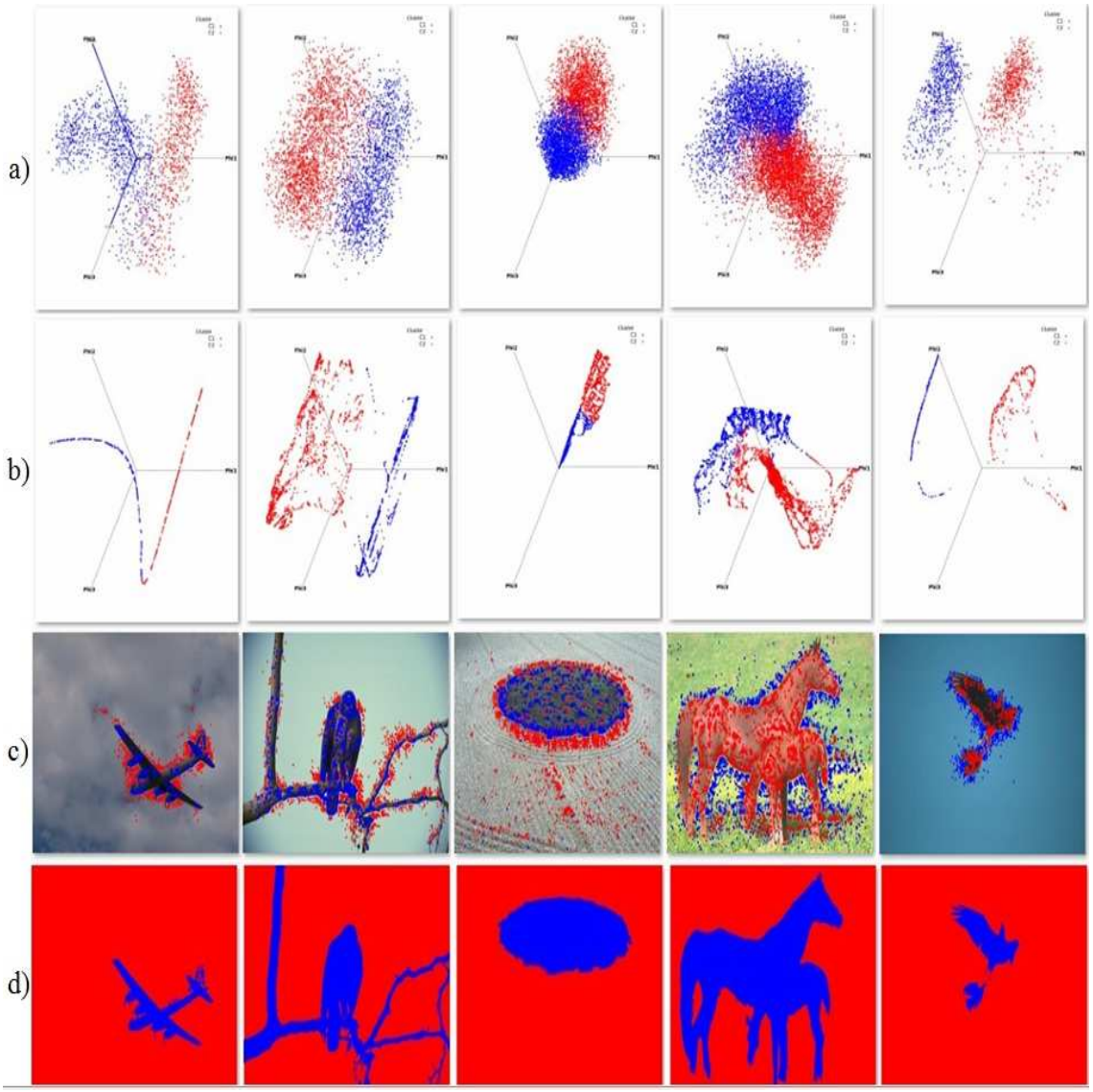

FIGURE 5.2 – Classification des points SIFT et leurs projections sur l'espace réduit et les images à segmenter

La figure 5.2 montre quelques résultats de classification des nuages des points SIFT projetés dans l'espace réduit, ainsi que leurs projections sur les images à segmenter.

Les graphes de la figure 5.2-(b) représentent les versions régularisées de ceux de la figure 5.2-(a), avec $p = 2$. Comme on peut le voir, ceci a permis de bien classer les germes de l'objet d'intérêt et ceux de l'arrière plan (figure 5.2-(c)). La figure 5.2-(d) montre le résultat de la propagation de labels (de la figure 5.2-(c)) sur le graphe associé à l'image initiale. Le résultat final est une segmentation binaire (cas de 2 classes : objet/fond). Ces résultats sont obtenus après 100 itérations en utilisant l'équation 5.2 .

Les F-measure, Rappel et Précision, calculés comme décrit dans [Kulkarni, 2009], et

correspondant à ces images, sont présentés dans la Table suivante (image1, image2, image3, image4 et image5 correspondent, respectivement, aux images dans la figure (5.2-(c)) de gauche vers la droite).

	image1	image2	image3	image4	image5
F-measure	0.8669	0.9211	0.9743	0.5937	0.9024
Rappel	0.9911	0.9718	0.9734	0.4532	0.9996
Précision	0.7703	0.8754	0.9751	0.8606	0.8225

TABLE 5.1 – F-measure, rappel et précision

Un autre exemple des images multi-label est présenté dans la figure 5.3. Cet exemple montre l'identification des points SIFT (figure (5.3-(a)) où les classes de ces derniers sont nettement visibles en utilisant la régularisation discrète sur graphe avec $p = 2$ (figure (5.3-(b)). Les projections de ces classes sur les images sont présentées dans les figures (5.3-(c) et 5.3-(d)).

Pour évaluer la performance de notre approche, nous avons utilisé deux mesures objectives pour évaluer la qualité de la segmentation : l'index de Rand (RI) [Rand, 1971] et l'erreur de cohérence globale (the global consistency error : GCE) [Hanbury, 2008]. Le RI mesure la cohérence d'un étiquetage entre une segmentation et sa vérité terrain par le rapport entre des paires de pixels ayant les mêmes étiquettes. L'objectif est d'attribuer deux pixels à la même classe si et seulement s'ils sont similaires et cela afin de mesurer le pourcentage de décisions qui sont corrects. Le GCE mesure à quel point une segmentation peut être considérée comme un raffinement de l'autre. Pour l'évaluation, nous notons que la mesure de similarité RI est meilleure quand elle est plus élevée et la mesure de distance GCE est meilleure quand elle est inférieure. Nous présentons les résultats qualitatifs et quantitatifs pour notre algorithme de segmentation et de trois autres, à savoir : l'algorithme Fuzzy C-Means (FCM) [Bezdek, 1984], l'algorithme WaterShed (WS) [Vincent, 1991]. et l'algorithme Mean-Shift (MS) [Comaniciu, 2002]. Nous avons implémenté ces algorithmes en utilisant la bibliothèque Pandore [Pandore].

La figure(5.4) présente les résultats qualitatifs de notre algorithme appliqué sur les mêmes images que la figure (5.2). On peut observé que lorsque les germes sont bien répartis entre les objets, la segmentation est très similaire à celle segmentation effectuée par un humain.

Pour avoir des résultats quantitatifs, des expériences ont été menées sur 100 images en utilisant les algorithmes ci-dessus. Souvent, l'évaluation GCE favorise la sur-segmentation.

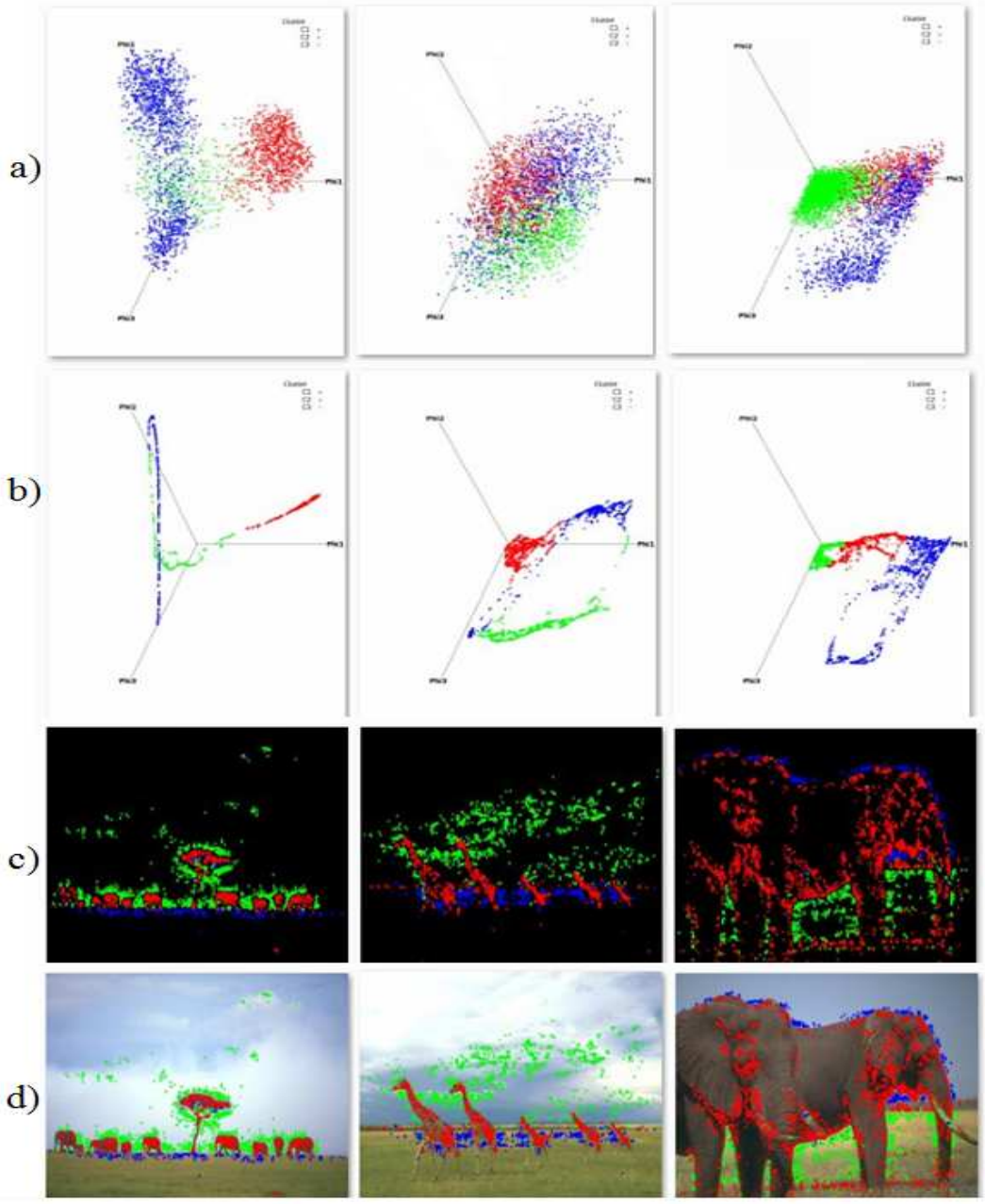

FIGURE 5.3 – Segmentation multi-label des images

Par conséquent, pour comparer nos résultats avec d'autres méthodes, nous avons procédé à la segmentation sans tenir compte des régions d'une superficie inférieure à 2% de l'image. Si la valeur de RI est plus élevé et la GCE est plus faible, alors l'approche de segmentation est meilleure.

La performance de notre approche par rapport à ces algorithmes est rapportée dans la Table 5.2. Comme nous pouvons le voir, notre méthode donne de meilleurs résultats. Elle a la plus faible mesure de la GCE et la plus grande mesure de la RI. La Table 5.2

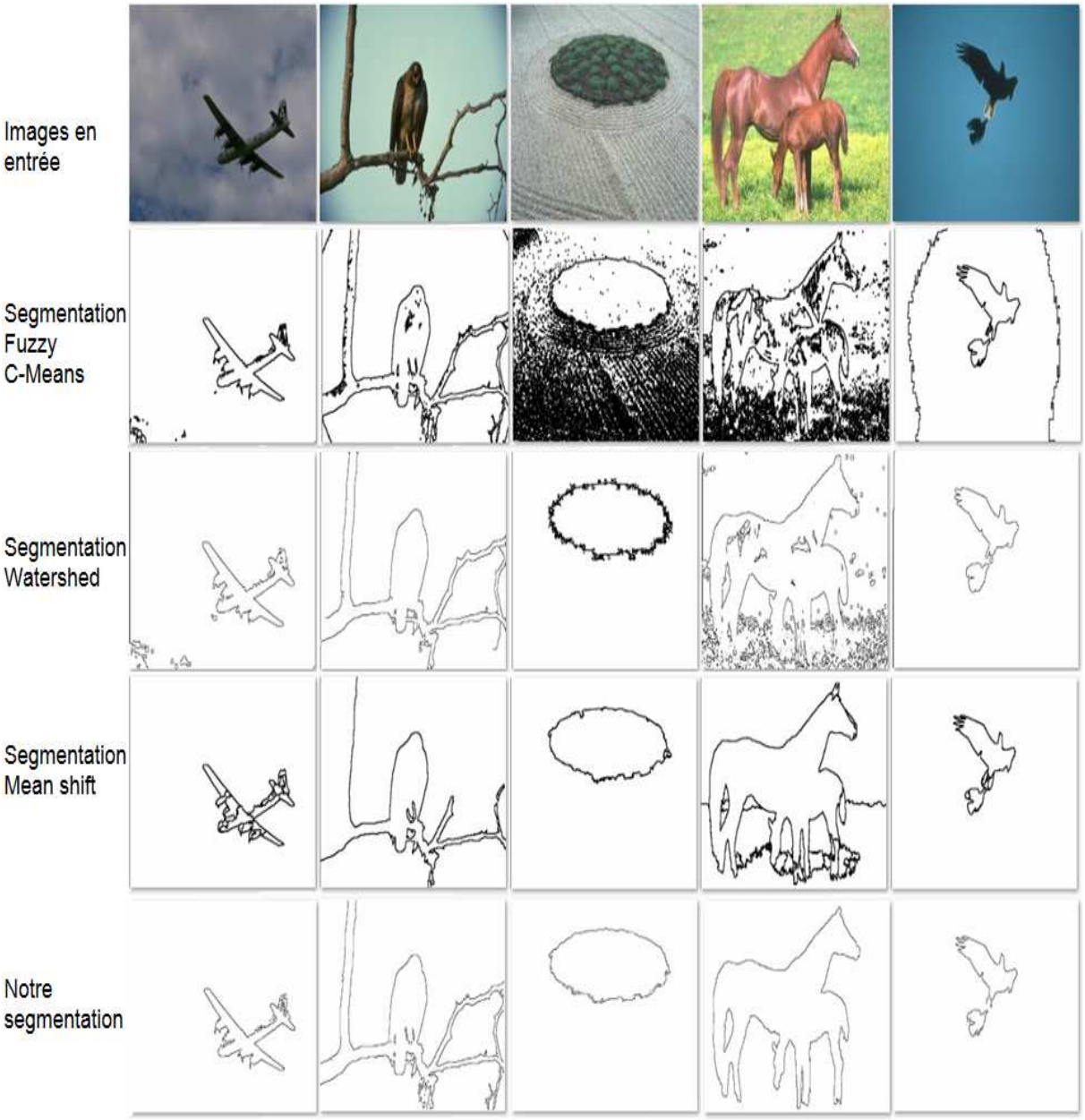

FIGURE 5.4 – Les résultats qualitatifs de notre approche

révèle également que l'indice de Rands (RI) de *Fuzzy c-means* est plus élevé que les autres algorithmes et l'erreur de cohérence globale de l'algorithme du *Watershed* est inférieure aux autres.

Les résultats expérimentaux montrent que notre approche peut segmenter les régions d'intérêt avec plus de précision que les autres méthodes de segmentation.

	GCE	RI
Vérité terrain (Humain)	0.079	0.875
Fuzzy C-Means	0.221	0.789
Watershed	0.203	0.697
Mean-Shift	0.259	0.755
Notre approche	**0.189**	**0.799**

TABLE 5.2 – L'évaluation des performances de notre algorithme

5.6 Conclusion

Dans ce chapitre, nous avons abordé le problème de la réduction-diffusion basées sur des graphes de données de grande dimension. Nous l'avons présenté sur le cas de la segmentation d'images en couleur où nous avons décrit un cadre unifié composé de trois phases :

1. une analyse spectrale des données de grande dimension pour en réduire la dimensionnalité en utilisant un graphe représentant les points SIFT,
2. un débruitage et une classification de ces points SIFT dans l'espace réduit avec p-laplacien et
3. une diffusion multi-labels des points caractéristiques dans le graphe initial des données de grande dimension.

Avec ce schéma, un ensemble de points-clés est automatiquement identifié sur l'image et bien réparti entre le fond et les objets d'intérêt (OI). Par la suite, ces germes sont propagés progressivement sur le graphe, représentant l'image, en exploitant les informations structurelles et visuelles des pixels et ce jusqu'à la segmentation de l'OI. Nous avons implémenté l'approche proposée et obtenu des résultats expérimentaux encourageants.

Chapitre 6

Structuration des flux TV en utilisant un framework de diffusion

Sommaire

Résumé :
Nous présentons dans ce chapitre, un cadre unifié pour l'analyse structurelle des flux TV en utilisant les marches aléatoires sur graphe d'états transition associé à une chaîne de Markov spatio-temporelle. Notre contribution comporte quatre volets. Tout comme dans le chapitre 4 et le chapitre 5, nous utilisons les propriétés spectrales du graphe de similarité pour déduire de nouvelles coordonnées euclidiennes et de faible dimension, correspondant aux caractéristiques extraites des différentes vidéos. Ensuite, nous régularisons le dernier graphe pour améliorer la catégorisation du flux TV.
Le troisième volet de notre contribution permet de construire une hiérarchie des résumés du flux TV. En effet, dans un premier temps, les classes identifiées dans le nouvel espace réduit sont considérées comme des états. Par conséquent, le flux TV peut être interprété soit comme une chaîne de caractères, ce qui permet de détecter les éventuelles répétitions, soit comme un graphe d'états transition. Dans ce dernier cas, un niveau est alors créé dans la hiérarchie des résumés du flux TV. Regrouper, encore une fois, les classes constituant ce niveau permet de créer un autre niveau de résumé plus générale. Ainsi, répéter ce processus de regroupement permet de créer différents niveaux de résumés allant du plus détaillé au plus général. Ces niveaux constituent une hiérarchie de résumés du flux TV où le spectateur peut composer son propre programme en ignorant les résumés qui ne l'intéressent pas ou en naviguant en profondeur dans d'autres résumés. Ceci se traduit par la réalisation d'une coupe multi-échelles dans la hiérarchie des résumés.
Enfin, dans le quatrième volet de notre contribution, nous décrivons notre algorithme de segmentation vidéo interactive multi-label qui propose au spectateur de retrouver dans son programme, les résumés qui partagent le même contenu visuel. Les résultats expérimentaux ont montré que notre framework permet de structurer le contenu vidéo et d'obtenir un taux de classification significatif.

6.1 Introduction

De nos jours, de très grandes bases de données vidéo sont disponibles. La plupart d'entre elles sont issues de chaînes de télévision et sont de différentes natures. En général, elles ont un certain degré d'homogénéité (par exemple, des journaux télévisés, des films, des émissions de sport, etc). Certaines catégories sont diffusées à des périodes bien connues et conservent leurs créneaux horaires tout au long de la semaine, ce qui génère un taux considérable de redondance du contenu visuel. D'autres flux TV sont fortement typés et ont un taux élevé de dissimilarités entre eux (par exemple, filmé en/hors studios), tandis que d'autres se caractérisent par un usage intensif de gros plans.

Comment classer automatiquement ces flux est un sujet important en vision par ordinateur. Une solution à ce problème facilitera, non seulement les applications telles que la recherche de vidéos par le contenu ou les systèmes de surveillance vidéo, mais améliore, aussi, la navigation automatique et permet un accès rapide à la matière visuelle requise.

6.2 État de l'art

Au bas de la structure vidéo, les trames d'images consécutives sont regroupées en plans. Chaque plan peut être représenté par une ou quelques images clés. Un rappel de la terminologie, fréquemment utilisée, est présenté dans [Xingquan, 2004]. De nombreuses techniques ont été proposées pour structurer les documents vidéo :

- *Les approches de détection des limites des plans :* Les premières études visent à découper les vidéos volumineuses en de plus petits morceaux [Truong, 2000,Hanjalic, 2002,Ayache, 2005,Ahmad, 2007] (voir, par exemple [Ahmad, 2007] pour un aperçu général des méthodes d'évaluation des algorithmes de segmentation de plans). Parmi les algorithmes classiques, nous pouvons citer celui de la différence des histogrammes de couleurs utilisé pour détecter les coupures durs (hard cuts), l'algorithme utilisant l'écart type des intensités des pixels pour détecter les changements de plans fondus (fade cuts) et l'algorithme basé sur l'utilisation du contraste des contours pour les changements doux de plans (des changements qui prennent un certain temps). Cependant, le problème difficile est de distinguer les limites des plans des mouvements de caméra et des changements d'illumination.
- *Les approches de détection des groupes :* [Bertini, 2001, Yeung, 2001, Piriou, 2004] L'objectif est de trouver une structuration de haut niveau en regroupant, principalement, les plans présentant un voisinage sémantique. Malgré que cette tâche

est difficile, lorsqu'on traite des corpus hétérogène, les scènes peuvent être détectées à partir des vidéos comme les journaux télévisés, les émissions de sport ou les bulletins de la météo, parce qu'elles obéissent à des critères subjectifs [Kender, 1998]. Pour ce faire, la similarité entre deux plans est calculée, soit en utilisant la distance entre les images représentées par leurs histogrammes d'intensité, soit en utilisant une distance entre deux ensembles de trames d'images.

– *Les approches d'extraction de programmes :* Un programme est considéré comme une émission de télévision régulière comme par exemple : les journaux télévisés, les bulletins météo, les talk-show, ou les émissions de sport. En comparaison avec les approches précédentes, peu d'études existent pour la détection des limites de programmes dans une diffusion TV. Nous ne pouvons distinguer que deux types d'approches dans ce domaine :

  1. celles basées sur la répétition afin de détecter les inter programmes et les séquences audiovisuelles quasi-identiques dans le flux TV [Naturel, 2008, Gauch, 2006, Herley, 2006, Berrani, 2008]. En effet, la plupart des inter programmes sont diffusés à plusieurs reprises, et
  2. celles basées sur la détection et qui utilisent les caractéristiques intrinsèques des inter-programmes comme les changements de sons ou la présence de logos [Lienhart , 1997, Albiol, 2004].

  Naturel et al. [Naturel, 2007] proposent une structuration rapide des flux TV à l'aide des guides de programmes pour étiqueter les émissions détectées. La méthode repose sur la détection des segments non-programme (comme les pauses commerciales). Liang et al. [Liang, 2005] supposent que pour les vidéos TV, les programmes apparaissent et se terminent à des moments relativement fixes tous les jours. Poli et al. [Poli, 2006] proposent de construire un modèle de prédiction par apprentissage, sur des données concernant la télévision, en utilisant les chaînes de Markov cachées.

6.3 Solution proposée

6.3.1 Les grandes lignes de la solution proposée

Dans notre cas, et pour permettre une analyse automatique des flux TV volumineux, nous proposons d'utiliser une structure hiérarchique comportant plusieurs niveaux de résumés allant du plus fin au plus grossier. A chaque niveau de résumé, des structures compactes partagent les mêmes caractéristiques visuelles. La structure proposée fournit divers résumés de la vidéo et offre, ainsi, un accès non linéaire au contenu visuel

désiré. Pour ce faire, d'abord les vidéos, composant le flux TV, sont découpées en plans et chaque plan est représenté par une images clé (keyframe). Ensuite, un graphe de similarité visuelle est construit sur ces keyframes. Chaque keyframe est décrite par les attributs présentés dans 2.2.1.3. Le spectre de ce graphe est analysé comme décrit dans le chapitre 2 afin de générer de nouvelles coordonnées euclidiennes de faible dimension. Puis, un autre graphe est construit sur ces nouvelles coordonnées et régularisé comme décrit dans le chapitre 3. Une fois les vidéos classées dans l'espace réduit régularisé, chaque keyframe est alors marquée par sa classe d'appartenance.

Ainsi, deux services peuvent être proposés à l'utilisateur : le premier est intuitif et consiste à détecter les régularités dans le flux vidéo (par exemple, des scènes répétées). Le deuxième service fournit des résumés proposés à différents niveaux d'abstraction, ce qui permet à l'utilisateur de personnaliser le programme qu'il souhaite regarder.

6.3.2 Détection des régularités par appariement de chaînes de caractères

Dans cette section, nous formulons le problème de la détection de la régularité comme un problème de filtrage. L'idée principale est de considérer les séquences vidéo, représentées par leurs images clés, comme un ensemble de chaînes de caractères où, initialement, chaque plan de la vidéo est décrit par une image clé. La séquence vidéo est ensuite transformée en une séquence de symboles à partir d'un alphabet défini. Chaque symbole dans la chaîne représente une classe, et l'ensemble des symboles forme l'alphabet. Par conséquent, le nombre total de symboles dépend des classes utilisées. Donc, vérifier si des séquences consécutives (comme un clip vidéo) sont répétées dans la vidéo revient à chercher la plus grande sous-séquence qui se répète constamment dans la chaîne de caractères représentant la vidéo.

L'algorithme 1 est utilisé pour détecter et compter les répétitions dans une séquence vidéo. L'objectif est de trouver les plus grandes sous-séquences qui se répètent (au-delà d'un seuil) dans la chaîne vidéo.

L'objectif, ici, est d'exploiter la répétition dans la chaîne vidéo pour améliorer le regroupement des plans. En particulier, trouver l'ensemble de plans consécutifs qui apparaissent à différents intervalles avec une certaine cardinalité.

Algorithm 1 algorithme de détection de régularités dans une vidéo

ENTRÉES :

$VS = S_0 ... S_N$: chaîne Vidéo (Video String).

m : une classe (un symbole de l'alphabet VS).

δ : un seuil pour le nombre de plans consécutifs. Au-dessous duquel, la séquence trouvée ne sera pas considérée comme répétitive.

τ : une durée de temps. Au-dessous duquelle, la séquence trouvée ne sera pas considérée comme répétitive.

SORTIES : k plus longues sous-séquences LS_m^k

1: $k \leftarrow 0$
2: $j \leftarrow INDEX(m, VS, 0)$ // retourne l'index du premier S_m dans VS à partir de la position 0.
3: **Tant que** la chaîne VS n'est pas entièrement parcourue **Faire**
4: $(VS_m, taille) \leftarrow FIND(S_m, VS, j)$ // retourne la chaîne suivante des "S_m" ainsi que sa taille à partir de la position j.
5: $j \leftarrow j + taille$
6: **Si** $(taille \geq \delta)$ et $(durée\_frames(VS_m) \geq \tau)$ **Alors**
7: $LS_m^k \leftarrow VS_m$
8: $k \leftarrow k + 1$
9: **fin Si**
10: **Fin Tant que**
11: **return** $\{ LS_m^0, \ldots, LS_m^{k-1} \}$

6.3.3 Génération de programmes personnalisés

Pour permettre à l'utilisateur de composer son propre programme, nous produisons un résumé hiérarchique du document vidéo et nous le représentons par un dendrogramme (voir la figure 6.1).

Formellement, soit $\mathscr{X} = \{x_1, x_2, \ldots, x_n\}$ l'ensemble fini représentant la hiérarchie d'une vidéo. Une hiérarchie $\mathbb{H}$ de $\mathscr{X}$ peut être définie par :

1. $\mathfrak{B}(\mathscr{X}) = \{x_i | x_i \subseteq \mathscr{X}\} \in \mathbb{H}$ (l'ensemble des partitions de $\mathscr{X}$)

2. $\forall A, B \in \mathbb{H}, A \cap B \in \{A, B, \varnothing\}$.
   c-à-d. deux classes sont soit disjointes ou contenues l'une dans l'autre.

Un niveau de $\mathbb{H}$ est construit en tenant compte des informations obtenues en résolvant le problème de marches aléatoires sur le graphe de transition construit sur les classes du niveau juste en dessous.

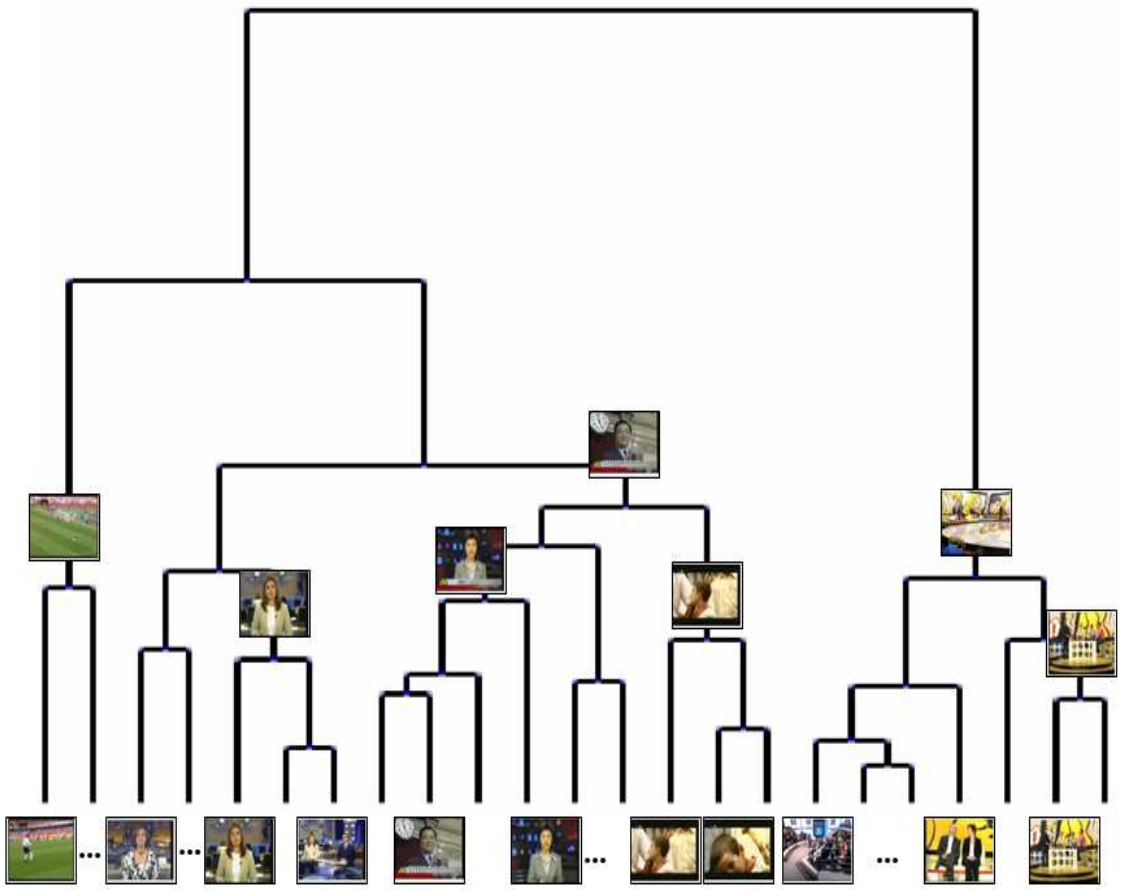

FIGURE 6.1 – Représentation hiérarchique d'un flux TV en utilisant un dendrogramme

6.3.3.1 Marches aléatoires sur graphe d'état-transition d'une chaîne de Markov spatio-temporelle

Le graphe construit sur les classes, appartenant à un niveau de la hiérarchie, peut être interprété comme un graphe d'états-transition d'une chaîne de Markov. Chaque état de cette chaîne représente une classe. Cette dernière peut être représentée par un key-frame. Utiliser les propriétés spectrales de ce graphe permet d'obtenir une classification plus générale qui fusionne les classes similaires dans des méta-classes. Ainsi, répéter ce processus de regroupement permet de créer différents niveaux de résumés allant du plus détaillé au plus général constituant, par conséquent, une hiérarchie de résumés du flux TV. Notons que le niveau le plus bas de cette hiérarchie représente les classes identifiées dans l'espace réduit régularisé. La figure 6.2 illustre ce principe.

A chaque niveau $l \in \mathbb{H}$, un graphe $G_l = (V_l, E_l, Dtw_l)$ est défini, où V_l est l'ensemble des classes et Dtw_l (Dynamic Time wrapping) mesure le degré de similarité entre chaque deux séquences vidéos et permet de calculer les probabilités de transition d'une classe à une autre.

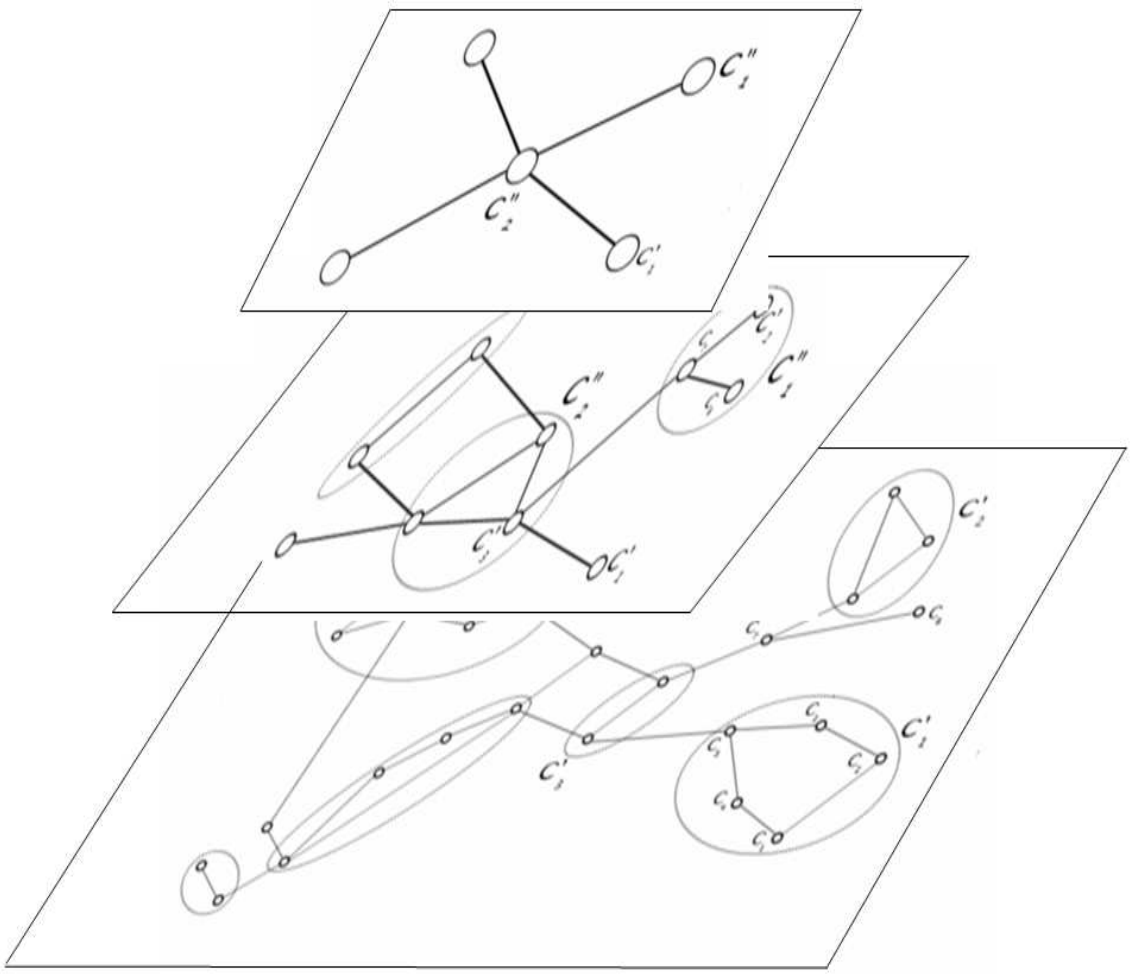

FIGURE 6.2 – Graphe d'état-transition pour la représentation des classes d'une hiérarchie

Étant donné que, dans notre framework, une vidéo est représentée par un ensemble de keyframes qui sont corrélées du point de vue temporel, l'approche basée sur DTW est plus appropriée pour la mesure des distances. Pour comparer deux chaînes vidéo (Video String) ($VS1$ et $VS2$) de différentes longueurs ($|VS1| < |VS2|$), nous évitons la comparaison par paires en mettant à l'échelle le plus long Video String $VS2$ au plus petit $VS1$ (garder un nombre égale au plus petit ensemble).

Soient, par exemple, R et Q deux vidéo string. DTW trouve un chemin optimal entre R et Q en utilisant une programmation dynamique pour calculer la distance minimale cumulée $\gamma(M,N)$, où $\gamma(i,j)$ est définie récursivement comme suit : $\gamma(i,j) = d(r_i,q_j) + \min(\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1))$ où $d(r_i,q_j)$ représente la distance entre le point de données r_i ($1 \leqslant i \leqslant M$) de la sequence R et le point de données q_j ($1 \leqslant j \leqslant N$) de la séquence Q. Nous avons utilisé l'approche proposée dans [Xie, 2010] pour définir cette distance.

6.3.3.2 Coupe Multi-echelle pour la composition de programmes

Pour composer un programme spécifique, l'utilisateur doit sélectionner les classes qui correspondent à son intérêt et cela aux différents niveaux de la structure hiérarchique (voir la figure 6.3).

Les classes sont disposées suivant la chronologie d'apparition de leurs keyframes représentatives. Ainsi, nous pouvons étendre la définition précédente de $\mathbb{H}$ pour encoder la position spatiale des classes dans la hiérarchie. La fonction g définissant ces coordonnées est donnée par :

$$\begin{aligned} g: \quad & \mathbb{H} \longrightarrow \mathbb{N}^2 \\ & x_l \longmapsto g(x_l) = (i, j) \end{aligned}$$

i indique l'ordre chronologique des classes et j la position de son niveau dans $\mathbb{H}$.

Formellement, un programme est considéré comme un sous-ensemble $\mathscr{C} \subsetneq \mathbb{H}$ composé des coordonnées des classes sélectionnées. Il est défini comme suit :

$$\mathscr{C}(\mathbb{H}) = \{g(x_l) | x_l \in \mathbb{H}\} \tag{6.1}$$

6.3.4 Segmentation Interactive multi-label des vidéos

Notre algorithme basé sur la p-régularisation est directement applicable à la segmentation semi supervisée. L'idée de base de notre méthode de propagation de labels est de considérer un algorithme itératif où chaque sommet absorbe certaines informations des labels de son voisinage et met à jour son propre label. La procédure peut être répétée jusqu'à ce que tous les sommets du graphe soient labellisés et ne changent plus.

Supposons qu'on a c classes, alors l'ensemble des étiquettes devient $L = \{1 \dots c\}$. Soit f une fonction qui assigne des labels pour chaque sommet du graphe. Initialement :

$$f^0(u) = \begin{cases} 0 & \text{if } u \text{ est non labellisé} \\ k = C_u(v) & k \in L, \text{ pour chaque sommet labellisé } v \sim u \end{cases} \tag{6.2}$$

$C_u(v)$ est une fonction de décision qui classe le sommet u en fonction de son voisinage labellisé. Si $k = 1$, le problème est ramené à une classification binaire et $L = \{0, 1\}$.

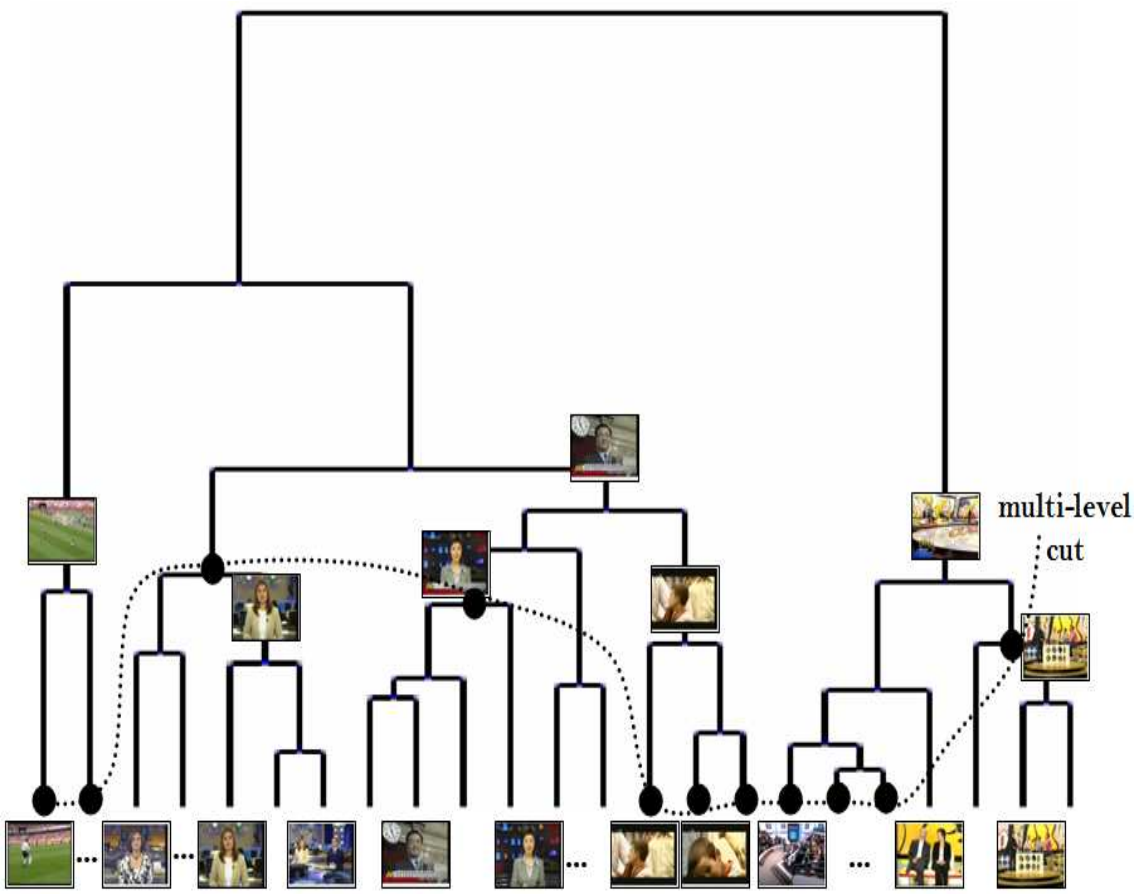

FIGURE 6.3 – Coupe multi-niveaux

Dans ce travail, nous traiterons le cas de deux classes. Un ensemble de sommets à étiquetés conformément aux souhaits de l'utilisateur (un label) et les autres sommets. Ainsi, Le principe de segmentation binaire, par propagation d'étiquettes, peut être décomposé en trois étapes, qui sont :

étape 1 : Choisissez deux classes de germes, une marquant le premier plan ($f(F) = 1$, les séquences qui intéressent l'utilisateur) et les autres marquent le background ($f(B) = 0$, les séquences qui n'intéressent pas l'utilisateur).

étape 2 : Trouver un optimum de l'équation (3.17)

étape 3 : Diffuser les germes initiaux

$$f(u) = \begin{cases} 1 & \text{if } f(u) \geqslant \text{ seuil, ou} \\ 0 = C_u(v) & \text{if } f(u) < \text{ seuil} \end{cases} \tag{6.3}$$

Par exemple, on peut choisir la valeur 0.5 pour le seuil (seuil = 0.5).

6.4 Expérimentations

– Base de données Video

Pour effectuer nos tests, nous avons utilisé un corpus vidéo qui est la propriété du Laboratoire INRIA (Institut National de Recherche en Informatique et en Automatique - France). Cette base de données est composée de différents types de flux de chaînes de télévision. Les images clés ont été déjà calculées et sont caractérisées par les descripteurs présentés dans la sous-section 2.2.1.3.

Les tests que nous avons effectués se sont concentrés sur le débruitage et la réduction de la dimensionnalité de ces flux vidéo. Dans un premier temps, nous utilisons les descripteurs des keyframes pour calculer les similarités visuelles afin de définir un noyau qui permet de définir des coordonnées dans l'espace réduit. Différentes distances sont utilisées, entre autres : la distance de Bhattacharya, la distance de kolmogorov-smirnov, la distance intersection et la distances de corrélation. L'algorithme de classification par k-means nous a permis de regrouper les keyframes en huit classes. Les résultats de la classification sont donnés dans la Table 6.1.

Classe	Kolmogorov	Bhattacharyya	Intersection	Correlation
1	55%	85%	98%	60%
2	83.35%	83.35%	83.35%	66.65%
3	75%	62.5%	66.7%	66.65%
4	56.25%	18.75%	62.5%	68.75%
5	71.85%	81.25%	83.4%	75%
6	84.4%	93.75 %	100%	90.6%
7	96.9%	90.6 %	96.9%	96.9%
8	84.4 %	71.9%	87.5%	75%

TABLE 6.1 – Taux de reconnaissance en utilisant différentes distances

La figure 6.4-a présente un aperçu de la projection des différentes vidéos dans l'espace réduit. La figure 6.4-b présente, quant à elle, le résultat de la régularisation du graphe avec $p = 2$. Comme nous pouvons le voir, la régularisation permet d'avoir des structures plus compactes. Il est également démontré que le processus de classification peut être facilement réalisé et donne de meilleurs résultats en particulier avec des valeurs décroissantes de p.

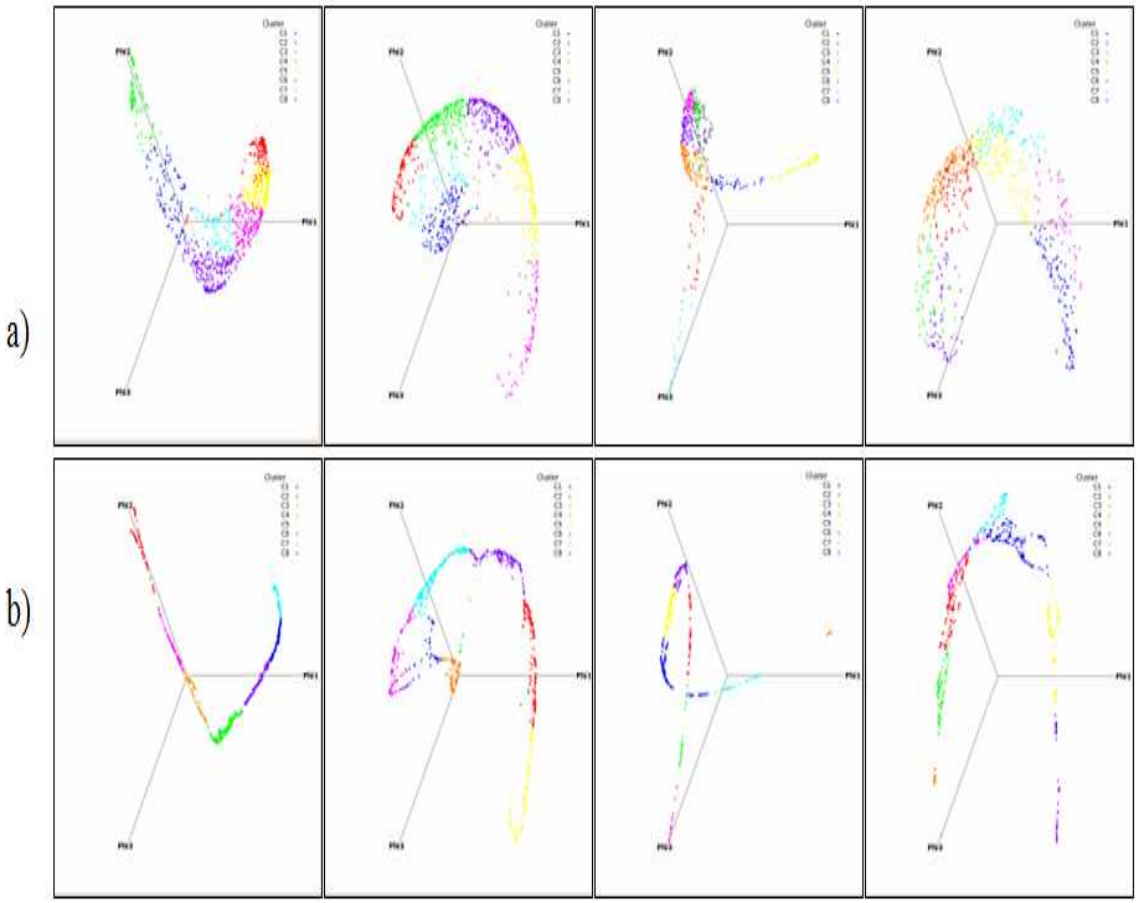

FIGURE 6.4 – Régularisation de variété avec p=2

6.5 Conclusion

Dans ce chapitre, nous avons abordé le problème de la structuration des flux TV en utilisant un cadre de diffusion. Nous avons présenté un framework unifié à travers trois étapes : décomposition spectrale pour la définition d'un espace réduit, débruitage de variété avec le p-laplacien et construction interactive de programmes. Le corpus vidéo utilisé contient divers programmes regroupant des journaux télévisés, des films, des émissions de sport et des spectacles de variétés.

Nous avons proposé une description du programme a travers un schéma de classification hiérarchique. La stratégie proposée tient compte de la hiérarchie de contenu et de la redondance entre les vidéos. Avec ce schéma, le programme vidéo peut être exprimé progressivement, de bas en haut, allant des résumés longs à des résumés plus généraux. Ceci permet de construire un graphe personnel qui exploite les informations sémantiques et les caractéristiques visuelles acquises des données vidéos.

Conclusion et perspectives

Les informations multimédias traitées par les applications informatiques deviennent de plus en plus volumineuses. La recherche d'une information particulière est devenue une tâche difficile qui prend beaucoup de temps. Cette difficulté est due principalement à la grande dimension des données multimédias qui sont, souvent, décrites par des vecteurs de taille importante. Ceci qui constitue un obstacle pour la recherche de similarité qui est à la base des algorithmes de classification et qui constitue par conséquent, un critère déterminant pour la validité des résultats obtenus. La recherche de similarité est directement liée à la recherche des plus proches voisins, or il a été démontré que la distance d'un objet à son plus proche voisin et sa distance à son voisin le plus éloigné ont tendance à converger lorsque la dimension de ces objets est très grande.

Différentes techniques visant à réduire la dimension des données ont été proposées. Leur but est la recherche des représentations appropriées en identifiant les dimensions qui sont porteuses de l'information pertinente tout en gardant l'erreur de reconstruction la plus proche de zéro afin de garder le maximum d'informations sur la structure originale des données.

L'utilisation des graphes pour le regroupement et la réduction de la dimension des données est relativement récente. En plus des propriétés statistiques des données, un graphe permet de saisir aussi leur géométrie, ce qui permet d'avoir des informations sur le voisinage local et non local.

L'objectif de cette thèse était la proposition d'un modèle adapté au données de grande dimension permettant une restructuration adéquate de la base de données en entrée en vue d'accélérer la recherche et y faciliter la navigation.

Le travail que nous avons réalisé rentre dans le cadre de l'analyse et l'organisation des données de grande dimension. L'idée de base de la solution que nous avons proposée consiste à considérer un graphe modélisant tous les objets d'intérêt (image, document,

etc.) d'une base à travers leurs descripteurs. Les sommets représentent les objets d'intérêt et la pondération des arêtes mesure la similarité entre ces objets. Cette approche permet :

– d'associer aux données multimédias, une description conservant leur cohérence présente dans la séquence à traiter.
– d'extraire un code contenant l'information pertinente.

Moyennant les marches aléatoires sur le graphe modélisant les similarités visuelles entre les données, nous en déduisons de nouvelles coordonnées réduites et euclidiennes ainsi qu'une nouvelle famille de distances appelées "diffusion maps".

Nos contributions

Notre travail a été structuré en deux parties. Dans la première partie nous avons défini les bases théoriques de notre modèle d'organisation des données de grande dimension. Dans la deuxième partie, nous avons validé notre approche sur les trois applications suivantes :

1. Un cadre unifié pour la reconnaissance des comportements humains.
2. Apprentissage faiblement supervisé à partir des points SIFT.
3. Structuration des flux TV en utilisant un framework de diffusion.

Nos contributions sont les suivantes :

1. Utilisation des moments de zernike 3D pour décrire la silhouette et la dynamique d'un objet d'intérêt dans un clip vidéo.
2. Définition d'un framework unifié d'organisation et de diffusion des informations sur graphe. Ce framework est composé des six modules suivants :
   (a) Module de prétraitement.
   (b) Module de calcul de similarité.
   (c) Module des marches aléatoires sur graphe.
   (d) Module d'accélération des marches aléatoires sur graphe.
   (e) Module de visualisation et classification.
   (f) Module de régularisation discrète sur graphe. Ce dernier module est constitué, à son tour, des deux sous-modules suivants :
      i. Sous-module d'élimination des données aberrantes.
      ii. Sous-module de prédiction des données manquantes.

3. Accélération des marches aléatoires sur graphe.

Des publications ont validé notre modèle adapté aux données de grande dimension.

Perspectives

Le présent travail est bien adapté pour le traitement des données déjà disponibles. Dans certains contextes, les informations sont acquises progressivement et la taille de la base de données n'est donc pas connue à l'avance. C'est le cas par exemple, des traitements en ligne. Dans de telles situations, nous pouvons utiliser des solveurs itératifs comme par exemple, la méthode Nyström. Cette méthode est utilisée pour réaliser une classification incrémentale et extrapoler, ainsi, celle déjà existante.

Cela permettra, non seulement, d'éviter le stockage exhaustif des données, mais aussi d'assurer une réaction rapide du système offrant, ainsi, une solution à un large spectre applicatif tels que l'aide à la navigation d'engins autonomes, l'aide à l'analyse et l'identification pour des systèmes de surveillance et d'observation, etc.

Par ailleurs, nous sommes entrain d'utiliser différentes modalités dans le cadre de notre projet PNR "identification multimodale du locuteur en temps réel" en intégrant d'autres caractéristiques audiovisuelles afin d'améliorer les résultats d'identification.

Bibliographie

[Ahmad, 2006] Mohiuddin. Ahmad and Seong-Whan. Lee. Hmm-based human action recognition using multiview image sequences. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 1, pages 263–266. IEEE Computer Society, 2006.

[Ahmad, 2007] Ashraf M. A. Ahmad. Multimedia content and the semantic web : Methods, standards and tools : Book reviews. *J. Am. Soc. Inf. Sci. Technol.*, 58(3) :457–458, 2007.

[Albiol, 2004] Alberto Albiol, M. J. Fulla, Antonio Albiol, and L. Torres. Detection of tv commercials. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 541–544, 2004.

[Anjum, 2001] Ali. Anjum and J. K. Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and Recognition of Events in Video'01.*, pages 28–28. IEEE Computer Society, 2001.

[Antonio, 2002] Antonio Robles-kelly, Sudeep Sarkar, and Edwin R. Hancock. A fast leading eigenvector approximation for segmentation and grouping. In *In Proc. 16th International 110 Conference on Pattern Recognition*, pages 639–6422. IEEE Computer Society Press, 2002.

[Ayache, 2005] Stéphane Ayache, Georges Quénot, and Shin'ichi Satoh. CLIPS and NII at TRECvid : Shot segmentation and feature extraction. In *TREC Workshop on Video Retrieval Evaluation*, 2005.

[Baker, 1977] Christopher T. H. Baker. *The numerical treatment of integral equations*. Monographs on numerical analysis. Clarendon Press, Oxford, New York, 1977.

[Ballan, 2009] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. In *Proceedings of the 16th IEEE international conference on Image processing*, pages 3533–3536. IEEE Press, 2009.

[Bay, 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3) :346–359, 2008.

[Bellman, 1961] Richard E. Bellman. *Adaptive control processes - A guided tour*. Princeton University Press, 1961.

[Belkin, 2003] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6) :1373–1396, 2003.

[Bertini, 2001] M. Bertini, A. Del Bimbo, and P. Pala. Content-based indexing and retrieval of tv news. *Pattern Recogn. Lett.*, 22(5) :503–516, 2001.

[Berrani, 2008] Sid-Ahmed Berrani, Gaël Manson, and Patrick Lechat. A non-supervised approach for repeated sequence detection in tv broadcast streams. *Image Commun.*, 23(7) :525–537, 2008.

[Beyer, 1999] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful ? In *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, pages 217–235. Springer-Verlag, 1999.

[Bezdek, 1984] J. Bezdek, R. Ehrlich, and W. Full. Fcm : The fuzzy c-means clustering algorithm. *Computers & Geosciences*, pages 191–203, 1984.

[Blake, 2004] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *in ECCV*, pages 428–441, 2004.

[Blank, 2005] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1395–1402. IEEE Computer Society, 2005.

[Bobick, 2001] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23 :257–267, 2001.

[Boyce, 1983] J.F. Boyce and W.J. Hossack. Moment invariants for pattern recognition. *Pattern Recognition Letters*, 1 :451–456, 1983.

[Boykov, 2001] Y Y Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. volume 1, pages 105–112, 2001.

[Boykov, 2003] Y Boykov and V Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *ICCV, 2003*, pages 26–33, 2003.

[Boykov, 2004] Y Boykov and V Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 :1124–1137, 2004.

[Bretzner, 2002] L. Bretzner, I. Laptev, and T. Lindberg. Hand gesture recognition using multi-scale color features, hierarchical models and particle filtering. *IEEE International Conf. on Automatic Face and Gesture Recognition*, pages 423–428, 2002.

[Campbell, 2010] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Automatic 3d object segmentation in multiple views using volumetric graph-cuts. *Image Vision Comput.*, 28(1) :14–25, 2010.

[Canterakis, 1999] N. Canterakis. 3d zernike moments and zernike affine invariants for 3d image analysis and recognition. In *In 11th Scandinavian Conf. on Image Analysis, 1999*, pages 85–93, 1999.

[Cattell, 1966] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, pages 245–276, 1966.

[Chen, 2007] Q Chen, N.D. Georganas, and E.M. Petriu. Real-time vision based hand gesture recognition using haar-like features. *IEEE Transactions on Instrumentation and Measurement*, pages 1–6, 2007.

[Coifman, 2006] R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1) :5–30, 2006.

[Comaniciu, 2002] Dorin Comaniciu and Peter Meer. Mean shift : A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5) :603–619, 2002.

[Costantini, 2011] L. Costantini, L. Seidenari, G. Serra, L. Capodiferro, and A. Del Bimbo. Space-time zernike moments and pyramid kernel descriptors for action classification. In *ICIAP (2)*, pages 199–208. Springer, 2011.

[Dhillon, 2009] P.S. Dhillon, S. Nowozin, and C.H. Lampert. Combining appearance and motion for human action classification in videos. *Computer Vision and Pattern Recognition Workshop*, pages 22–29, 2009.

[Efros, 2003] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pages 726–733. IEEE Computer Society, 2003.

[Fujiyoshi, 1998] Hironobu. Fujiyoshi and Alan J. Lipton. Real-time human motion analysis by image skeletonization. In *In Proceedings of IEEE WACV98*, pages 15–21. IEEE Computer Society, 1998.

[Gauch, 2006] John M. Gauch and Abhishek Shivadas. Finding and identifying unknown commercials using repeated video sequence detection. *Comput. Vis. Image Underst.*, 103(1) :80–88, 2006.

[Hanjalic, 2002] Alan Hanjalic. Shot-boundary detection : Unraveled and resolved ? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2) :90–105, 2002.

[Hanbury, 2008] Allan Hanbury and Julian Stoettinger. On segmentation evaluation metrics and region counts. In *ICPR*, pages 1–4, 2008.

[Harris, 1988] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[Henrik, 2008] Carl Henrik. Ek, Philip H. S. Torr, and Neil D. Lawrence. Gaussian process latent variable models for human pose estimation. In *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, pages 132–143. Springer-Verlag, 2008.

[Herley, 2006] C. Herley. Argos : automatically extracting repeating objects from multimedia streams. *Trans. Multi.*, 8(1) :115–129, 2006.

[Hotelling, 1933] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24 :417–441, 1933.

[Huang, 2008] Ling Huang, Donghui Yan, Michael I. Jordan, and Nina Taft. Spectral clustering with perturbed data. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, pages 705–712. Curran Associates, Inc., 2008.

[Karhunen, 1946] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae*, 34 :7, 1946.

[Ke, 2004] Yan Ke and Rahul Sukthankar. Pca-sift : a more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition*, pages 506–513. IEEE Computer Society, 2004.

[Kellokumpu, 2011] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Recognition of human actions using texture descriptors. *Machine Vision and Applications*, 22 :767–780, 2011.

[Kender, 1998] John R. Kender and Boon lock Yeo. Video scene segmentation via continuous video coherence. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 367–373, 1998.

[Kim, 2008] Jong-Sung Kim and Ki-Sang Hong. A new graph cut-based multiple active contour algorithm without initial contours and seed points. *Mach. Vision Appl.*, 19(3) :181–193, 2008.

[Kolmogorov, 2004] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 :65–81, 2004.

[Kulkarni, 2009] F. Nicolls M. Kulkarni. Interactive image segmentation using graph cuts. In *Twentieth Annual Symposium of the Pattern Recognition Association of South Africa*, 2009.

[Lafon, 2006] Stephane Lafon, Yosi Keller, and Ronald R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11) :1784–1797, 2006.

[Lafon, 2006] Stephane Lafon and Ann B. Lee. Diffusion maps and coarse-graining : A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9) :1393–1403, 2006.

[Laptev, 2004] C. Schuldt, I. Laptev, and Barbara. Caputo. Recognizing human actions : A local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 32–36. IEEE Computer Society, 2004.

[Lawrence, 2003] Lawrence. N. Gaussian process latent variable models for visualisation of high dimensional data. In *In NIPS, 2003*, 2003.

[Lawrence, 2005] N. Lawrence and A. Hyvarinen. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6 :1783–1816, 2005.

[Lederman, 1987] S.J. Lederman and R.L. R.L. Klatzky. Hand movements : A window into haptic object recognition. *Cognitive Psychology*, 19(3) :342–368, 1987.

[Li, 2003] Ren-Cang Li. On perturbations of matrix pencils with real spectra, a revisit. *Math. Comput.*, 72(242) :715–728, 2003.

[Liang, 2005] Liuhong Liang, Hong Lu, Xiangyang Xue, and Yap-Peng Tan. Program segmentation for tv videos. In *International Symposium on Circuits and Systems (ISCAS 2005)*, pages 1549–1552. IEEE, 2005.

[Lienhart , 1997] R. Lienhart, C. Kuhmuench, and W. Effelsberg. On the detection and recognition of television commercials. In *Proc. International Conference on Multimedia Computing and Systems*, pages 509–516, 1997.

[Loeve, 1948] Loeve M. Fonctions aleatoires du second ordre. *Processus stochastiques et mouvement Brownien*, 1948.

[Lowe, 2004] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2) :91–110, 2004.

[Martin, 2001] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *in Proc. 8th Int'l Conf. Computer Vision*, pages 416–423, 2001.

[Mikolajczyk, 2005] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10) :1615–1630, 2005.

[Morel, 2009] Jean-Michel Morel and Guoshen Yu. Asift : A new framework for fully affine invariant image comparison. *SIAM J. Img. Sci.*, 2(2) :438–469, 2009.

[Nadler, 2005] Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *in Advances in Neural Information Processing Systems 18*, pages 955–962. MIT Press, 2005.

[Naturel, 2007] Xavier Naturel, Guillaume Gravier, and Patrick Gros. Fast structuring of large television streams using program guides. In *Proceedings of the 4th international conference on Adaptive multimedia retrieval : user, context, and feedback*, pages 222–231. Springer-Verlag, 2007.

[Naturel, 2008] Xavier Naturel and Patrick Gros. Detecting repeats for video structuring. *Multimedia Tools Appl.*, 38(2) :233–252, 2008.

[Novotni, 2004] Marcin Novotni and Reinhard Klein. Shape retrieval using 3d zernike descriptors. *Computer Aided Design*, 36 :1047–1062, 2004.

[Nyström, 1930] E. J. Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54(1) :185–204, 1930.

[Pandore] http ://www.greyc.ensicaen.fr/ regis/pandore/.

[Pavlovic, 1997] V.I. Pavlovic, R. Sharma, and T.S. Huang. Visual interpretation of hand getures for human-computer interaction. a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) :677–695, 1997.

[Pearson, 1901] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572, 1901.

[Piriou, 2004] Gwenaelle Piriou, Patrick Bouthemy, and Jian feng Yao. Extraction of semantic dynamic content from videos with probabilistic motion models. In *8th Eur. Conf. on Comp. Vis., ECCV'04*, pages 145–157, 2004.

[Poli, 2006] Jean-Philippe Poli and Jean Carrive. Modeling television schedules for television stream structuring. In *Proceedings of the 13th international conference on Multimedia Modeling*, pages 680–689. Springer-Verlag, 2006.

[Rand, 1971] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336) :846–850, 1971.

[Rosales, 2001] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3d hand pose reconstruction using specialized mappings. *IEEE International Con. on Computer Vision*, pages 378–385, 2001.

[Rother, 2004] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut" : interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3) :309–314, 2004.

[Roweis, 2000] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 :2323–2326, 2000.

[Sinop, 2007] Ali K. Sinop and Leo Grady. A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm. pages 1–8. IEEE, 2007.

[Stewart, 1990] G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory (Computer Science and Scientific Computing)*. Academic Press, 1990.

[Teh, 1988] Cho-Huak. Teh and Roland T. Chin. On image analysis by the methods of moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10 :496–513, 1988.

[Tenenbaum, 1998] Joshua B. Tenenbaum. Mapping a manifold of perceptual observations. In *NIPS '97 : Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 682–688. MIT Press, 1998.

[Tenenbaum, 2000] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 :2319–2323, 2000.

[Truong, 2000] Ba Tu Truong, Chitra Dorai, and Svetha Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 219–227. ACM, 2000.

[Urtasun, 2006] R. Urtasun. 3d people tracking with gaussian process dynamical models. In *In CVPR*, pages 238–245. CVPR, 2006.

[Vezzani, 2010] R. Vezzani, D. Baltieri, and R. Cucchiara. Hmm based action recognition with projection histogram features. In *Proceedings of the 20th International conference on Recognizing patterns in signals, speech, images, and videos*, pages 286–293. Springer-Verlag, 2010.

[Vincent, 1991] Luc Vincent and Pierre Soille. Watersheds in digital spaces : An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(6) :583–598, 1991.

[Wang, 2007] Jue Wang, Maneesh Agrawala, and Michael F. Cohen. Soft scissors : an interactive tool for real time high quality matting. *ACM Trans. Graph.*, 26(3), 2007.

[Wang, 2008] Jack M. Wang, David J. Fleet, and Aaron. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30 :283–298, 2008.

[Williams, 2001] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

[Wu, 2001] Y. Wu, J. Y. Lin, and T. S. Huang. Capturing natural hand articulation. *In Proc. 8th Int. Conf. on Computer Vision*, 2 :426–432, 2001.

[Xie, 2010] Bryan Wiltgen Ying Xie. Adaptive feature based dynamic time warping. In *IJCSNS*, volume 10, pages 264–273, 2010.

[Xingquan, 2004] Xingquan Zhu, Xindong Wu, Jianping Fan, Ahmed K. Elmagarmid, and Walid G. Aref. Exploring video content structure for hierarchical summarization. *Multimedia Syst.*, 10(2) :98–115, 2004.

[Xinghua, 2009] Sun. Xinghua, Chen. Mingyu, and A. Hauptmann. Action recognition via local descriptors and holistic features. *Computer Vision and Pattern Recognition Workshop*, pages 58–65, 2009.

[Xu, 2007] Ning Xu, Narendra Ahuja, and Ravi Bansal. Object segmentation using graph cuts based active contours. *Comput. Vis. Image Underst.*, 107(3) :210–224, 2007.

[Yeung, 2001] Minerva Yeung, Boon-Lock Yeo, and Bede Liu. Extracting story units from long programs for video browsing and navigation. pages 360–369. Morgan Kaufmann Publishers Inc., 2001.

[Zelnik, 2001] L. Zelnik-manor and M. Irani. Event-based analysis of video. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 123–130, 2001.

[Zhu, 2009] Guangyu. Zhu, Ming. Yang, Kai. Yu, Wei. Xu, and Yihong. Gong. Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 165–174. ACM, 2009.