Contributions à la caractérisation & structuration d'images et de séquences vidéo

Youssef Chahir

université de Caen
Basse-Normandie

UNIVERSITÉ de CAEN BASSE-NORMANDIE

U.F.R. Sciences

ÉCOLE DOCTORALE SIMEM

MÉMOIRE

présenté par

Mr Youssef CHAHIR

en vue de l'obtention de l'

Habilitation à Diriger des Recherches

Spécialité : Informatique et applications

Contributions à la Caractérisation & Structuration d'images et de séquences vidéo

soutenue publiquement le 27 Avril 2012
devant le jury composé de :

Rapporteurs	Pr Atilla Baskurt , LIRIS - INSA, Lyon
	Pr Patrick Lambert, LISTIC - Polytech Annecy-Chambéry , Annecy
	Pr Jack-Gérard Postaire, LAGIS - USTL, Lille
Examinateurs	Pr Liming Chen, LIRIS, Ecole Centrale de Lyon, Lyon
	Pr Abderrahim Elmoataz, GREYC - Université de Caen
	Pr François Jouen, CHART- EPHE, Paris
	Pr Florence Sèdes, IRIT, l'Université Paul Sabatier, Toulouse

Remerciements

Je tiens tout d'abord à remercier Atilla Baskurt, Patrick Lambert et Jack-Gérard Postaire qui ont accepté d'être rapporteurs de cette Habilitation à Diriger des Recherches et y consacrer du temps malgré leurs nombreuses obligations. Je les remercie pour leur lecture attentive et avisée de ce manuscrit.

J'adresse également mes remerciements à Florence Sèdes, François Jouen et Liming Chen pour avoir accepté d'être membres de ce jury et pour leur soutien et leur amitié. Que toutes ces personnes reçoivent ici l'expression de ma profonde gratitude et ma haute considération.

Un merci tout particulier à Abderrahim Elmoataz. Son aide et son soutien le long de ces années m'ont été sources de grandes satisfactions, tant professionnelles que personnelles.

Bien évidemment, je remercie aussi tous les étudiants, doctorants et autres chercheurs qui m'ont fait le plaisir de travailler avec moi, et sans l'aide desquels cette habilitation n'aurait pas de signification.

Je remercie également tous les membres du GREYC en particulier les équipes DoDoLa et Image pour toutes les années passées auprès d'eux.

Enfin, je ne pourrais terminer ces remerciements sans m'adresser à ma famille (A.M.I.S.) à qui je dédie ce mémoire en les remerciant pour leur patience et leur soutien indéfectible.

Table des matières

1

Introduction générale

Sommaire

1.1 Parcours

J'ai débuté mes activités de recherches en 1996 en optimisation avec contraintes au Laboratoire HEUDIASYC UMR CNRS 6599 à Compiègne sous la responsabilité de Pierre Villon. Ensuite, j'ai continué en thèse en indexation multimédia jusqu'en 2000 au sein de ICTT & MAPLY (UMR 5585) à l'Ecole Centrale de Lyon sous la direction de Liming Chen. Après une année d'ATER (1998 - 1999) au sein du département de Mathématiques de l'Ecole Centrale de Lyon, j'ai été nommé Maître de Conférences en Informatique à l'Université de Caen en Septembre 2000 et rattaché au laboratoire GREYC (Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen).

A mon arrivée, j'ai été intégré au sein de l'équipe de **fouille de données** *DoDoLa* (Données, Document, Langue) jusqu'à 2006 où j'ai travaillé sur le Document Electronique Composite (DEC) [1] composé principalement de textes et d'images. On s'intéressait particulièrement au document visuel et aux rapports existants entre les parties textes et images. Cette période comprenait des activités en fouille de données et apprentissage notamment lors des travaux sur l'indexation qui nécessite l'utilisation de classification automatique ou de découvertes de régularités dans les documents. En janvier 2007, j'ai rejoint l'équipe ***Image*** du GREYC où je travaille sur l'Extraction et Gestion des Connaissances (EGC) [2] dans les bases d'images et de vidéos et particulièrement sur l'analyse d'événements en vidéo.

Les travaux que je présente dans ce mémoire d'habilitation à diriger des recherches correspondent à des recherches effectuées au sein des équipes "data-mining" et image du GREYC depuis septembre 2000. Globalement, ceci est le fruit de travaux que j'ai réalisés avec des étudiants de master, des doctorants, ainsi qu'avec des collègues chercheurs.

1. DEC : un thème de l'équipe DoDoLa du GREYC.
2. EGC : un thème de l'équipe Image du GREYC.

1.2 Contexte de recherche

Les évolutions et progrès technologiques récents ont donné naissance à de nouvelles applications multimédias, qui se caractérisent par l'accroissement important du volume des données à acquérir, stocker puis traiter et par la distribution et l'hétérogénéité de ces données. L'analyse de ces données nécessite un large spectre de compétences allant de l'amélioration de l'image à sa reconstruction, de la reconnaissance et de l'extraction de formes et d'objets à l'indexation des données.

Mes travaux ont été, bien sûr, très influencés par les chercheurs que j'ai côtoyés depuis ma thèse et par l'évolution des domaines de recherche. Le premier problème auquel j'ai essayé d'apporter quelques contributions concerne l'extraction de nouveaux descripteurs pour la caractérisation d'images et de vidéos. Le deuxième problème concerne la création d'outils de base pour le traitement et l'analyse des données en vidéo, avec des projets interdisciplinaires liés à l'analyse des gestes et du regard. La motivation conjointe à ces études scientifiques réside dans la volonté de proposer des idées et/ou des outils permettant de réaliser une analyse de flux d'images et de vidéos afin d'accéder à des informations importantes sur les objets contenus dans les scènes observées que ce soit à des fins de détection, de classification ou de reconnaissance.

1.3 Organisation du document

Ce mémoire est composé de deux parties principales qui peuvent se lire indépendamment l'une de l'autre.

La première partie est consacrée au traitement de séquences d'images par modèles discrets sur graphes. Je présente des méthodes de régularisation discrètes pour la restauration de données bruitées et manquantes, et pour la catégorisation de séquences d'images et de vidéos. Je propose une approche de restauration non locale par patchs pour certains types de dégradations, et une autre manière de définir un ordre de données par une réduction non linéaire des dimensions grâce à des marches aléatoires sur graphes. Le domaine est vaste, et j'ai choisi d'organiser cette présentation en trois volets :

- Dans un premier temps, j'aborde la thématique de restauration de données bruitées en vidéo en exploitant la redondance spatio-temporelle et l'auto similarité et ce en utilisant un cadre de régularisation discrète sur graphes basé sur le p-Laplacien.
- Ensuite, je présente nos deux approches de restauration de données manquantes (inpainting). La première est basée sur la restauration par régularisation. Il s'agit d'une extension du modèle présenté en débruitage. La seconde approche est basée sur les fonctions p-harmoniques sur graphe qui constituent un cadre général d'interpolation permettant la réalisation de nombreuses applications dont l'inpainting.
- Enfin, je présente le dernier volet qui est la classification par une réduction non linéaire de dimension grâce à des marches aléatoires sur graphe. Je présenterai le lien avec la régularisation discrète et des applications en caractérisation de la saillance visuelle, en segmentation et en recherche par le contenu.

En seconde partie, je présente mes travaux sur l'analyse de séquences vidéo qui se proposent d'explorer les problématiques liées à l'analyse de gestes et d'actions d'une personne

en mouvement. Dans cette partie, je commence par présenter succinctement le processus de caractérisation de la couleur de peau à partir d'un espace hybride de couleurs. Dans la section suivante, je présente notre méthode de segmentation convexe d'un objet en mouvement à partir des contours actifs et le flot optique. Ensuite, j'aborde la classification de bases de *gestes et actions* en utilisant la régularisation discrète sur graphe et les moments spatio-temporels comme descripteurs des vidéos. Enfin, je présente un système de capture et de prédiction du regard par processus gaussiens. Le mémoire se conclut par une synthèse des principaux résultats obtenus et montre les perspectives de recherche ouvertes par ces travaux. En annexes, on trouvera en première partie une présentation de mon curriculum vitae qui retrace mon parcours, mes responsabilités pédagogiques, administratives, de recherche ainsi que mes projets en cours. En seconde partie, il y a la liste de mes publications.

2

Régularisation spatio-temporelle sur graphes

Sommaire

Résumé

Ce chapitre présente mes travaux basés sur les graphes pour la restauration de données bruitées (débruitage) et manquantes (inpainting) et pour la classification de données dans les séquence d'images. Les fondements sont basés sur des concepts provenant de la théorie des graphes et de l'analyse spectrale et exploitent les propriétés spectrales du Laplacien ainsi que les modèles de diffusion associés.

Le graphe permet de représenter des données discrètes de nature diverse et de capturer les liens entre elles. Deux problèmes majeurs constituent une entrave pour l'analyse et la classification de ces données : La dégradation des données (bruit, quantification, discrétisation, etc.) et la grande dimension des données. La majorité des traitements de ces données tels que le débruitage, l'interpolation ou l'organisation se résument à des problématiques d'analyse de graphes.

Ce chapitre présente les recherches que j'ai menées en régularisation de séquences d'images. Ces travaux ont été réalisés essentiellement avec Abderrahim Elmoataz dans le cadre de la thèse de M. Ghoniem et du master de K-E. Aziz que nous avons co-encadrés. D'autres travaux liés au processus de diffusion et à la caractérisation de séquences d'images ont été effectués dans le cadre de la thèse de A. Bouziane que j'encadre actuellement et de ma collaboration avec H. Tabout.

Les travaux présentés dans ce chapitre proposent d'adapter un modèle variationnel de régularisation en utilisant les graphes et leurs propriétés spectrales pour la résolution de problèmes liés à la restauration, la réduction de dimensions et la classification dans les séquences d'images. Deux méthodologies envisageables sont abordées pour exploiter la diffusion sur graphes et la classe de Laplaciens : une approche fonctionnelle basée sur les fonctions sur graphe et une autre approche matricielle où le filtrage est basé sur des matrices de Markov et qui correspond à un filtrage dans le domaine spectral.

C'est dans le cadre de la première approche que proposons de résoudre un certain nombre de problèmes inverses définis sur graphes, tels que le débruitage ou l'inpainting vidéo. Dans le cadre de la seconde approche, nous étudions les propriétés des marches aléatoires sur graphe (Matrice de transition) et nous présentons nos contributions pour des applications en réduction de dimension, en segmentation ou encore en apprentissage semi supervisé.

Le plan du chapitre est organisé comme suit :

- Dans la section 2.1, nous commençons par introduire les notions et les définitions issues de la théorie des graphes, en particulier une famille d'opérateurs, qui sont nécessaires à la bonne compréhension de la suite de ce manuscrit.
- Dans la section 2.2, après un rappel de la problématique et des grandes tendances avec les méthodes variationnelles existantes en débruitage vidéo, nous présentons notre algorithme de régularisation sur graphes basé sur le p-laplacien pour le débruitage des séquences d'images et la manière de construire le graphe.
- Dans la section 2.3, nous rappelons la problématique de l'inpainting et les grandes classes de méthodes proposées dans la littérature pour résoudre ce problème. Ensuite, nous présentons l'application de la régularisation itérative à des fins d'inpainting ainsi qu'une seconde approche d'interpolation basée sur les fonctions p-harmoniques.
- Dans la section 2.4, nous présentons une synthèse des méthodes spectrales pour la réduction de la dimension et la classification des images. Ensuite, nous formulons une approche de régularisation spectrale sur graphes basée sur les marches aléatoires sur graphe. Nous proposons un algorithme de segmentation à partir des points caractéristiques de l'image et nous présentons la détection des éléments saillants dans une images par marches aléatoires sur graphes.

2.1 Graphes et traitement de séquences d'images

Le graphe est une structure naturelle qui permet de représenter des données discrètes [New06]. La figure 2.1 montre quelques exemples d'utilisation de graphes. De nombreuses méthodes basées sur des graphes ont été proposées dans le domaine de l'analyse, du traitement et la classification de données. Elles sont basées sur des concepts provenant de la théorie des graphes et de l'analyse spectrale et exploitent les propriétés spectrales du Laplacien. Ces méthodes sont devenues très populaires pour de nombreuses applications telles que la réduction de dimension, le regroupement de données similaires ou la classification [LL06] [SM00] [BN08] [ELB08] [HM07].

Dans cette section je commence par rappeler de manière brève les notations et les définitions de la théorie des graphes. Ensuite, je présente une famille d'opérateurs, en particulier une classe de Laplaciens, définis sur des graphes [ELB08].

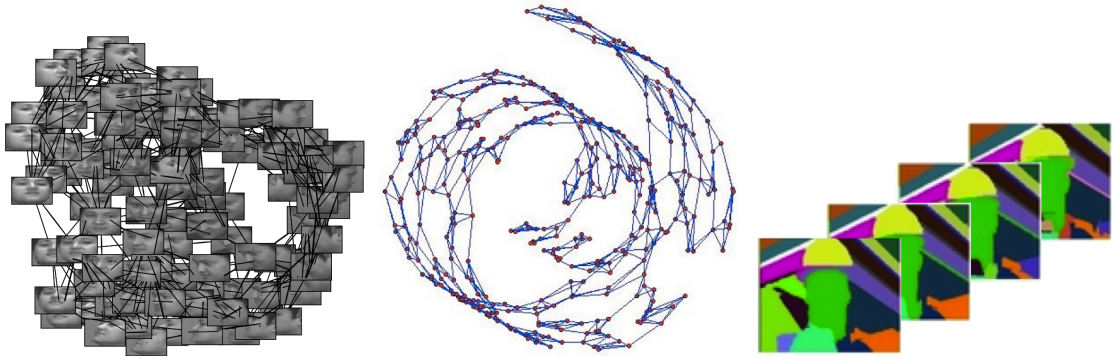


FIGURE 2.1: Exemples de graphes représentant des données diverses.

2.1.1 Définitions et notations

A partir d'un ensemble de N objets, il est possible d'obtenir une représentation détaillée de cet ensemble sous la forme d'un graphe pondéré noté $G = (V, E)$ où les nœuds représentent les objets présents et les arcs représentent les relations que l'on souhaite mettre en évidence. L'ensemble fini V désigne l'ensemble des sommets (noeuds) du graphe et E est un ensemble d'arêtes (arcs) inter-nœuds :

$$E = \{(u, v) \subseteq V \times V \text{ avec } u \neq v \text{ et } u, v \in V\}.$$

L'ordre $|G|$ d'un graphe est défini comme étant le nombre de sommets de G i.e. $|G| = |V(G)|$.

Deux noeuds u et v appartenant à V sont adjacents si l'arête $(u, v) \in E$. Les deux noeuds sont alors appelés des noeuds voisins. Le voisinage $\mathcal{N}(v)$ d'un sommet v désigne l'ensemble de ses voisins. Il est défini par :

$$\mathcal{N}(v) = \{u : \forall u \in V, (u, v) \in E\}.$$

Nous désignons par $u \sim v$ tout sommet u qui appartient au voisinage de v ($u \in \mathcal{N}(v)$).

On notera A **la matrice d'adjacence** du graphe avec $A(u, v) = 1$ si $u \sim v$, 0 sinon.

Un graphe est non-orienté quand l'ensemble des arêtes est symétrique, c'est à dire que pour chaque arête $(u, v) \in E$, nous avons également $(v, u) \in E$.

Une chaîne de G est une liste ordonnée de sommets de G telle que chaque sommet de la liste soit adjacent au suivant.

Un graphe est connexe si $\forall u \in V$ et $\forall v \in V$, il $\exists$ une chaine de u vers v c-à-d qu'il existe une suite d'arêtes permettant d'atteindre v à partir de u.

2.1.2 Fonctions définies sur des graphes

Soit $\mathcal{H}(V)$ l'espace de Hilbert des fonctions à valeurs réelles sur les noeuds du graphe G. Soit $f : V \to \mathbb{R}^n$ une fonction qui attribue à chaque noeud $v \in V$ un vecteur $f(v)$ de dimension n. Cette fonction peut être vue comme un vecteur colonne.

Par analogie avec l'espace des fonctions continues, l'intégrale discrète de la fonction f sur G est définie par :

$$\int_V f = \sum_{v \in V} f(v)$$

L'espace des fonctions $\mathcal{H}(V)$ est muni du produit scalaire :

$$\langle f, g \rangle_{\mathcal{H}(V)} = \sum_{v \in V} f(v)g(v) \quad \text{avec } f, g : V \to \mathbb{R}$$

De la même manière, soit $\mathcal{H}(E)$ l'espace des fonctions $F : E \to \mathbb{R}^+$définies sur les arêtes et également muni du produit scalaire :

$$\langle F, G \rangle_{\mathcal{H}(E)} = \sum_{(u,v) \in E} F(u, v)G(u, v) \quad \text{avec } F, G : E \to \mathbb{R}$$

La norme $\mathcal{L}_2$ [1] d'une fonction f obtenue à partir du produit scalaire est : $\|f\|_2 = \sqrt{\langle f, f \rangle_{\mathcal{H}(V)}}$.

Dans ce mémoire, nous ne considérons que des graphes simples (sans arêtes multiples et sans boucles) non-orientés et pondérés.

2.1.2.1 Fonction de poids

Un graphe est pondéré si une fonction de poids $w \in \mathcal{H}(E)$ lui est associée. Soit G=(V,E,w) où la fonction $w : E \to \mathbb{R}^+$ désigne la fonction de poids du graphe et satisfait les conditions suivantes :

$$\begin{cases} w(u, v) = w(v, u) & \forall (u, v) \in E \\ w(u, v) > 0 & si\ v \in \mathcal{N}(u) \\ w(u, v) = 0 & si\ v \notin \mathcal{N}(u) \end{cases}$$

1. Norme p :
$\mathcal{L}_p(f) = \|f\|_p = (|f_1|^p +) \cdots + |f_n|^p)^{\frac{1}{p}}$
$\mathcal{L}_\infty(f) = \|f\|_\infty = max(|f_1| + \cdots + |f_n|)$

Cette fonction de poids rend compte des similarités entre les données. Elle dépend généralement de l'application considérée et de la variété des données.

On notera W **la matrice des poids** $w(u,v)$ du graphe. Cette matrice est symétrique et semi-définie positive. Si le choix du poids est binaire, on retrouve $W = A$.

Le degré $\deg : V \to \mathbb{R}^+$ d'un noeud u est la somme des poids des arêtes incidentes à ce sommet :

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

On notera D **la matrice diagonale des degrés** $deg(v)$.

Le volume d'un ensemble de sommets $\mathcal{A} \subseteq V$ $Vol(A) : V \to \mathbb{R}^+$ est une mesure de la *taille* du sous-ensemble $\mathcal{A}$ et se définit par :

$$vol(\mathcal{A}) = \sum_{v \in A} deg(v)$$

2.1.2.1.a Vecteurs caractéristiques

Dans nos applications, les similarités entre les noeuds du graphe reposent sur une comparaison entre les vecteurs caractéristiques qui lui sont associés. Ces vecteurs dépendent généralement d'une fonction $f : V \subset \mathbb{R}^m \to \mathbb{R}^n$ définie en chaque noeud. m est la dimension de la grille d'observation et n est la dimension des données à traiter. Pour une séquence d'images, si on considère que nous avons un graphe de voxels v_i représentés par leurs couleurs en RGB, alors $m = 3$ (pour un voxel (x,y,t)) et $n = 3$ (pour la couleur (R,G,B).

On définit le **support d'un noeud** v noté $B(v) \subseteq \{v\} \cup \mathcal{N}(v)$ comme le noeud v avec l'ensemble de ses noeuds voisins qui sont impliqués dans la constitution du patch.

Le vecteur caractéristique $\mathcal{F}(f,v) \in \mathbb{R}^m$ en chaque sommet $v \in V$ du graphe dépendant de la fonction f est défini alors par :

$$\mathcal{F}(f,v) = (f(u) : u \in \mathcal{B}(v))^T \tag{2.1}$$

Un graphe peut être pondéré par une distance ou une mesure de similarité entre les noeuds. La similarité entre deux noeuds dépend généralement d'une mesure de distance $d : E \to \mathbb{R}^+$ entre les vecteurs caractéristiques.

Un exemple classique de fonction de similarité est :

$$w(u,v) = exp\left(-d(u,v)^2/2\sigma^2\right) \text{ avec } d(u,v) = d(\mathcal{F}(f,u), \mathcal{F}(f,v)) = \left\|\mathcal{F}(f,u) - \mathcal{F}(f,v)\right\|_2$$

σ est fixé a priori ou il peut être estimé par l'écart-type sur la fonction$f$.

Dans le cas le plus simple, où nous n'avons à traiter qu'un noeud v, nous avons $F(f,v) = f(v)$ et donc la distance Euclidienne est :

$$d(f(u), f(v)) = \sqrt{\sum_{i=0}^{n-1} (f_i(u) - f_i(v))^2}$$

2.1.2.1.b Graphe de voisinage

Il existe plusieurs façons de construire un graphe de voisinage (V, E, w) en fonction du type de voisinage que l'on utilise. Les voisinages les plus couramment utilisés sont :

- Le ε-voisinage ($\mathcal{N}_\varepsilon(v)$). Une arête est créée entre le sommet u et le sommet v si la distance entre u et v est inférieure à une valeur ε. Plusieurs mesures de la distance sont envisageables, la plus fréquente est la distance euclidienne. Le voisinage $\mathcal{N}_\varepsilon(v)$ est défini par :
$$\mathcal{N}_\varepsilon(v) = \{u : \forall u \in V, d(u, v) \leq \varepsilon\}.$$
La distance qui est souvent utilisée est : $d(u, v) = \left\|u - v\right\|_2^2$ pour contrôler la proximité spatio-temporelle entre les noeuds.
- Le k-plus proche voisinage. Une arête est créée entre un sommet u et un sommet v si v appartient au k-plus proche voisinage de u ou si u appartient au k-plus proche voisinage de v. L'ensemble des arêtes considérées sont :
$$E = \{(u, v) : u \in \mathcal{N}_\varepsilon(v) \; ou \; v \in \mathcal{N}_\varepsilon(u), \}.$$
- Le k-plus proche voisinage mutuel. Une arête est créée entre un sommet u et un sommet v si v appartient au k-plus proche voisinage de u et si u appartient au k-plus proche voisinage de v. Nous considérons l'ensemble des arêtes suivant :
$$E = \{(u, v) : u \in \mathcal{N}_\varepsilon(v) \; et \; v \in \mathcal{N}_\varepsilon(u), \}.$$

2.1.2.2 Propriétés spectrales

2.1.2.2.a Matrice Laplacienne

Le laplacien est un opérateur qui est directement lié à la théorie spectrale des graphes. Il est utilisé dans de nombreux traitements et applications allant du pré-traitement des données jusqu'à leur classification. On trouve dans la littérature différentes définitions du Laplacien. sous différentes expressions :

- Laplacien combinatoire :

$$L : L(u, v) = \begin{cases} deg(v), & \text{si } u = v \\ -w(u, v), & \text{si } u \sim v \\ 0, & sinon \end{cases}$$

- Laplacien normalisé :

$$L_n : L_n(u, v) = \begin{cases} 1 - \frac{w(u,v)}{deg(v)}, & \text{si } u = v \\ -\frac{w(u,v)}{\sqrt{deg(u)deg(v)}}, & \text{si } u \sim v \\ 0, & sinon \end{cases}$$

Remarques :

- $L = D - W$.
- $L_n = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$

Ces matrices sont symétriques, semi-définies positives et possèdent des valeurs propres positives. Une fonctionnelle de régularisation d'une fonction f, couramment utilisée et associée à la matrice Laplacienne L, est la suivante :

$$f^T L f = f^T (D - W) f = \sum_{u \sim v} w(u, v)(f(u) - f(v))^2 \tag{2.2}$$

La minimisation de cette dernière revient à calculer les vecteurs propres de la matrice $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ Cette fonctionnelle et ses variantes, les processus de diffusion qui y sont associés, ainsi que l'étude des propriétés spectrales des Laplaciens (valeurs et vecteurs propres), sont à la base de nombreuses méthodes d'analyse de données discrètes, comme par exemple la réduction de dimensions, le regroupement, la segmentation d'images, ou encore la classification semi-supervisée.

2.1.2.2.b Energie du graphe

Le concept de l'énergie du graphe $E(G)$ a été introduit par Uvan Gutman en 1978 [GZ06] comme étant la somme des valeurs propres de la matrice d'adjacence A :

$$E(G) = \sum_{i=0}^{|G|} \lambda_i \tag{2.3}$$

V. Nikiforov [Nik07] a généralisé ce concept et a proposé une définition de l'énergie d'une matrice $E(M)$ comme étant la somme de ses valeurs singulières s_i :

$$E(M) = \sum_{i=0}^{|G|} s_i(M) \tag{2.4}$$

L'énergie d'un graphe G selon sa matrice associée M est défini par :

$$E_M(G) = \sum_{i=0}^{|G|} |\lambda_i - \bar{\lambda}| \tag{2.5}$$

où λ_i sont les valeurs propres de la matrice M et $\bar{\lambda}$ la moyenne des valeurs propres.

2.1.2.3 Opérateurs de différence sur graphe

Dans cette section, nous rappelons succinctement les définitions des opérateurs différentiels sur graphes introduites par Elmoataz et al. [ELB08] [ELBT08][BEM09].

⋆ L'opérateur de différence pondérée $d_w : \mathcal{H}(V) \to \mathcal{H}(E)$ d'une fonction $f \in \mathcal{H}(V)$ est défini en un sommet u selon une arête $uv \in E$ par :

$$d_w f(uv) = \sqrt{w_{uv}}(f(v) - f(u)) \tag{2.6}$$

Par analogie avec le calcul différentiel en continu, nous pouvons voir la différence pondérée discrète définie sur un graphe appliquée à une arête $uv \in E$:

$$\left.\frac{\partial_f}{\partial_e}\right|_v = \partial_v f(u) = d_w f(uv) \tag{2.7}$$

⋆ L'opérateur de divergence sur graphe est l'opérateur $div : \mathcal{H}(V) \to \mathcal{H}(E)$ qui satisfait :

$$\langle df, g\rangle_{\mathcal{H}(E)} = \langle f, -div(g)\rangle_{\mathcal{H}(V)}$$

avec $f \in \mathcal{H}(V)$ et $g \in \mathcal{H}(E)$. L'opérateur de divergence sur graphe en un noeud u d'une fonction g est défini par :

$$(-div(g))(u) = \sum_{v \in \mathcal{N}(u))} \sqrt{w_{uv}}(g(u,v) - g(v,u)) \tag{2.8}$$

⋆ Le gradient pondéré $\nabla_w : \mathcal{H}(V) \to \mathbb{R}^m$ d'une fonction $f \in \mathcal{H}(V)$ en un sommet $u \in V$ est interprété comme la régularité de la fonction dans le voisinage du sommet. Plusieurs normes ont été proposées :

$$\left\|\nabla_w f(u)\right\|_p = \left(\sum_{v \in \mathcal{N}(u))} \left(\sqrt{w(u,v)}(f(v) - f(u))\right)^p\right)^{1/p} \quad \forall\ p \in]0, +\infty[\tag{2.9}$$

$$\left\|\nabla_w f(u)\right\|_\infty = \max_{v \sim u} \left(\sqrt{w(u,v)}|f(v) - f(u)|\right) \tag{2.10}$$

⋆ Les p-Laplaciens isotrope et anisotrope peuvent être obtenus à partir d'une formulation générale.

Soit $\Delta^*_{w,p,q} : \mathcal{H}(V) \to \mathcal{H}(V)$, le p-Laplacien pondéré d'une fonction f qui est défini par :

$\Delta^*_{w,p,q} f(u) =$

$$\frac{1}{2} \sum_{v \in \mathcal{N}(u))} w(u,v)^{q/2} \left(\left\|\nabla_w f(u)\right\|_q^{p-q} + \left\|\nabla_w f(v)\right\|_q^{p-q}\right).(f(u) - f(v))|f(u) - f(v)|^{q-2} \tag{2.11}$$

$\forall\ p, q \in]0, +\infty[$

Remarque :

Le p-Laplacien (2.11) est relié à différentes formulations de l'opérateur de Laplace. Dans le cas où le paramètre p = q = 2, le p-Laplacien correspond au Laplacien combinatoire [Chung, 1997]. Soit le p-Laplacien isotrope $\Delta^i_{w,p} = \Delta^*_{w,p,2}$ est défini par :

$$\Delta^i_{p,w} f(u) = \frac{1}{2} \sum_{v \in \mathcal{N}(u))} w(u,v) \left(\left\|\nabla_w f(u)\right\|_2^{p-2} + \left\|\nabla_w f(v)\right\|_2^{p-2}\right)(f(u) - f(v)) \tag{2.12}$$

Soit le p-Laplacien anisotrope $\Delta^a_{w,p} = \Delta^*_{w,p,p}$ qui est défini par :

$$\Delta^a_{p,w} f(u) = \sum_{v \in \mathcal{N}(u))} w(u,v)^{p/2} \left|f(u) - f(v)\right|^{p-2} (f(u) - f(v)) \tag{2.13}$$

2.2 Débruitage d'images par régularisation

☞ Publications associées : [RI6, CI10, CI11] [ThèseF1]

Cette section aborde la tâche de lissage pour la réduction du bruit dans les images et les vidéos. Il existe de nombreuses approches de lissage mais nous avons choisi de faire le parallèle avec les approches basées sur la régularisation continue par EDP uniquement car notre objectif est de transposer une de ces méthodes variationnelles au domaine discret sur graphe. Il s'agit du célèbre modèle ROF [ROF92b] reconnu pour son efficacité en débruitage. L'idée est de représenter les séquences d'images et de vidéos par un graphe et de traiter ce graphe. Notre contribution réside dans la mise en place d'algorithmes de régularisation sur graphes par p-Laplacien, avec $p \in]0, +\infty[$, pour le débruitage et la simplification des séquences d'images.

2.2.1 Introduction

Les images et films peuvent subir des dégradations de différentes natures durant l'acquisition, l'enregistrement et/ou la transmission. Ces défauts se traduisent souvent par des modifications locales des niveaux de gris (ou des couleurs) : bruits (additif gaussien, impulsionnel, poivre et sel), changements de contrastes locaux, etc. Il est important de rappeler que certaines dégradations sont propres aux séquences d'images vidéo [SBdHK11] telles que les dégradations temporelles qui incluent le flou de mouvement, le scintillement ou les artefacts temporels. Si la modélisation du problème de la restauration est commune à l'image et à la vidéo, les difficultés rencontrées ne sont pas les mêmes.

Dans cette section, nous nous intéressons au débruitage de données en vidéo. Il existe de nombreuses approches dans la littérature pour cette problématique qui proviennent des méthodes statistiques, de la théorie de l'information, des transformées dans le domaine fréquentiel, des EDPs et des méthodes variationnelles [SW00] [CS05a] [AK06] [Win06].

Par souci de cohérence, nous nous limitons aux méthodes variationnelles basées sur la régularisation continue par équations aux dérivées partielles (EDPs) car notre objectif est de transcrire et d'adapter certaines de ces EDP dans un cadre discret. La restauration d'un pixel dépend de la valeur de ses pixels voisins et la prise en compte du temps et du mouvement est un enjeu fondamental pour la restauration vidéo. On peut répartir les méthodes variationnelles existantes selon deux modèles de traitement [PE07] :
∗ *Traitement local basé sur l'estimation du mouvement* : C'est le modèle traditionnel qui prend en compte des contraintes du mouvement dans la séquences (contraintes spatiales et temporelles) autour d'un pixel. Pour estimer le mouvement, il est courant de faire une hypothèse d'illumination constante : Un point réel visible par la caméra aura la même couleur ou intensité de gris au cours du temps. Cette hypothèse se traduit donc par l'équation du flot optique [1]. Ce mouvement n'est pas toujours mesurable (ni visible par

1.
$$\frac{\partial I}{\partial t} + v \cdot \nabla I = 0 \tag{2.14}$$

l'œil). La première catégorie de méthodes se base sur l'estimation du mouvement et filtre les trajectoires [ZPP06]. La seconde compense le mouvement par une estimation du flot optique, puis filtre les images résultant de la compensation du mouvement [GAM+06].
∗*Traitement non local basé sur les patchs* : ce modèle se propose de supprimer les contraintes spatiales et temporelles locales et de les remplacer par des contraintes sur des patchs. Ce modèle tourne en sa faveur la difficulté du problème de l'ouverture en évitant l'estimation du mouvement tout en préservant des caractéristiques liées aux éléments répétitifs. Ce modèle est devenu populaire avec les moyennes non locales (NL-means), qui après avoir été mises en oeuvre pour le bruit gaussien par Buades et al [BCM08], ont connu de nombreuses améliorations (patchs de tailles variables) ou adaptations à d'autres types de bruits [DDT09], [DD10] [KB08] [BKB07] [DFE07].

La famille de filtres non locaux peut être considérée comme une régularisation basée sur des fonctionnelles non locales. Kindermann et al. [KOJ05] ont été les premiers à interpréter le filtre moyen non local et les filtres de voisinage comme une régularisation basée sur des fonctionnelles non locales. Plus tard, Gilboa et Osher [GO07a] ont proposé une fonctionnelle quadratique de différences pondérées pour la régularisation d'images. L'idée des méthodes de restauration par patchs est simple et séduisante à la fois, car elle repose sur la propriété que, dans les séquences d'images, un patch peut se retrouver presque à l'identique à plusieurs endroits de la séquence. Plusieurs travaux récents ont proposé un cadre de régularisation discrète avec une fonctionnelle d'énergie portant sur des données non locales [PMD+10] [ELB08] [ZS05]. C'est dans le cadre de l'approche non locale sur graphes que se situent nos travaux.

2.2.2 Problématique

Soit une fonction f définie sur un graphe qui représente une séquence d'images. On va supposer aussi tout au long de ce travail que le bruit qui affecte la séquence d'images peut être modélisé par une perturbation aléatoire de la séquence initiale. Dans un souci de simplicité, on fait l'hypothèse encore plus restrictive que le bruit est additif et gaussien.

Ainsi, si l'on observe une fonction $f^0 : V \subset \mathbb{R}^3 \to \mathbb{R}^3$, une version dégradée d'une fonction "originale" f, le modèle de bruit additif gaussien s'exprime alors de la façon suivante :

$$f^0(x) = f(x) + \eta(x), \tag{2.15}$$

où pour tout $x \in \Omega$, $\eta(x)$ est un vecteur gaussien de $\mathbb{R}^{d_c}$ (bruit gaussien à moyenne nulle et de variance σ^2).

La restauration de la fonction f^0 consiste à retrouver une estimation de f qui représente les données utiles. Une approche classique pour résoudre ce problème utilise la méthode de régularisation et consiste à minimiser la fonctionnelle à deux termes :

$$\mathcal{J}(f) = E_{lissage}(f) + \lambda E_{attache}(f^0, f) \tag{2.16}$$

$E_{lissage}$ est un terme de régularisation représentant l'information a priori sur la fonction à traiter. $E_{attache}$ est un terme de fidélité aux données. Le rapport de force entre ces deux termes est contrôlé par le multiplicateur de Lagrange λ. La minimisation de $\mathcal{J}$ revient alors à trouver la fonction f, suffisamment régulière sur Ω, tout en étant suffisamment

proche de la fonction f^0 observée. La solution de ce problème est généralement obtenue en considérant l'équation d'Euler-Lagrange associée à l'énergie $\mathcal{J}$.

Pour illustrer cette technique de régularisation, nous considérons un modèle variationnel basé sur la p-forme de Dirichlet qui étend la régularisation linéaire et la variation totale [ROF92b] :

$$f^* = \min_f \left\{ \frac{1}{p} \int_\Omega \left\|\nabla f\right\|_2^p + \frac{1}{2} \int_\Omega \lambda \left\|f^0 - f\right\|_2^2 \right\} \text{avec } p \in]0, +\infty[\tag{2.17}$$

où $\left\|\nabla f\right\|_2^p$ est la norme L_2 du gradient de la fonction f. Le premier terme est le terme de régularisation, et le deuxième terme est le terme d'approximation.

2.2.3 Formulation variationnelle

Étant donnés un graphe pondéré $G = (V, E, W)$ et une fonction $f^0 \in \mathcal{H}(V)$, le but est de trouver une fonction $f^* \in \mathcal{H}(V)$ qui est non seulement lisse sur G mais suffisamment proche de f^0. Nous formalisons la régularisation discrète p-Laplacienne d'une fonction $f^0 \in \mathcal{H}(V)$ par le problème suivant de minimisation sur graphe :

$$f^* = \min_{f \in \mathcal{H}(V)} \left\{ \tfrac{1}{p} \sum_{v \in V} \left\|\nabla f(v)\right\|^p + \tfrac{1}{2} \sum_{v \in V} \lambda \left\|f^0(v) - f(v)\right\|^2 \right\} \text{avec } p \in]0, +\infty[\tag{2.18}$$

L'équation d'Euler-Lagrange associée au problème de minimisation 2.18 s'écrit :

$$\Delta f^*(v) + \lambda(f^0(v) - f^*(v)) = 0 \quad \forall v \in V \tag{2.19}$$

En substituant l'expression du p-Laplacien, nous obtenons :

$$f^*(v) = \frac{\lambda f^0(v) + \sum_{u \sim v} \gamma_{f^*}(u, v) f^*(u)}{\lambda + \sum_{u \sim v} \gamma_{f^*}(u, v)} \quad \forall v \in V \tag{2.20}$$

avec $\gamma_f(u, v) = w(u, v)(\left\|\nabla f(v)\right\|^{p-2} + \left\|\nabla f(u)\right\|^{p-2})$

Les problèmes de minimisation peuvent être résolus par plusieurs méthodes numériques qui convergent efficacement vers la solution. Pour la régularisation spatio-temporelle sur les graphes pondérés, nous avons utilisé l'algorithme itératif de Gauss-Jacobi qui donne l'équation de diffusion discrète suivante :

Soit V l'ensemble des noeuds du graphe, à chaque itération $t+1$, $\forall\ v \in V$ et $\forall (u, v) \in E$

$$\left\{ \begin{array}{rcl} f^{(0)} & = & f^0 \\ \gamma_f^{(t)}(u, v) & = & w(u, v)(\left\|\nabla f^{(t)}(v)\right\|^{p-2} + \left\|\nabla f^{(t)}(u)\right\|^{p-2}) \\ f^{(t+1)}(v) & = & \frac{\lambda f^0(v) + \sum_{u \sim v} \gamma_f^{(t)}(u,v) f^{(t)}(u)}{\lambda + \sum_{u \sim v} \gamma_f^{(t)}(u,v)} \end{array} \right. \tag{2.21}$$

Le paramètre λ est une constante choisie a priori et la diffusion discrète se comporte comme un filtre moyenneur de voisinage qui s'adapte aux valeurs filtrées au cours des itérations.

2.2.4 Construction du graphe

Selon la topologie du graphe, et le choix de w, la régularisation peut être locale ou non-locale. La notion de non localité est prise en compte à travers :

- La connexité et le voisinage pour la recherche des candidats les plus similaires par rapport au noeud central.
- Le vecteur d'attributs pour comparer les candidats.

La non localité, telle qu'a été introduite par Buades et al. [BCM08], consiste à comparer un patch centré autour d'un pixel à l'ensemble des patchs existants dans l'image. Du fait de l'évidente complexité spatiale et temporelle de la méthode pour le traitement des images, la notion de non localité est usuellement remplacée en pratique par celle de semi localité. Cette dernière consiste à comparer, non plus l'ensemble des patchs de l'image pour un pixel donné, mais uniquement ceux se trouvant dans une fenêtre de recherche centrée sur ce dernier. Des travaux permettant de déterminer, par méthodes statistiques et de manière optimale et automatique, la taille des fenêtres de recherche et la taille des patchs ont été proposés. Il y a également des algorithmes qui cherchent à réduire la complexité des traitements non locaux des images [KB08] [DCC+08] [PGM10].

Nous considérons une séquence vidéo comme une fonction f définie sur les sommets d'un graphe pondéré $G = (V, E, w)$ où chaque sommet v est défini par un triplet (x_v, y_v, t_v) qui indique la position spatiotemporelle du pixel courant. Nous définissons le $(k_1 \times k_2 \times k_3)$-voisinage $\mathcal{N}(v)$ d'un sommet v par :

$$\mathcal{N}_{k_1,k_2,k_3}(v) = \left\{u = (x_u, y_u, t_u) \in V : \left|x_u - x_v\right| \leqslant k_1, \left|y_u - y_v\right| \leqslant k_2, \left|t_u - t_v\right| \leqslant k_3\right\} \tag{2.22}$$

De même, nous définissons un support 3D $\mathcal{B}(v)$ autour d'un noeud v comme une boîte de taille $r_x \times r_y \times r_t$. Il est intéressant de noter que r_x, r_y et r_t doivent être assez petits par rapport à k_1, k_2 et k_3 pour être sûr que le voisinage $\mathcal{N}_{k1,k2,k3}$ contienne un nombre significatif de patchs. Les relations suivantes doivent être respectées : $k_1 > \alpha_1 \times r_x$, $k_2 > \alpha_2 \times r_y$, et $k_3 \geq \alpha_3 \times r_t$. En pratique, nous utilisons par défaut $\alpha_1 = \alpha_2 = 3$ et $\alpha_3 = 3$. Nous définissons le vecteur caractéristique $\mathcal{F}(f, v) \in \mathbb{R}^m$ en chaque sommet $v \in V$ du graphe dépendant de la fonction f par :

$$\mathcal{F}(f, v) = (f(u) : u \in \mathcal{B}(v))^T \tag{2.23}$$

Ce vecteur peut incorporer des attributs variés décrivant un noeud du graphe tels que la couleur, la texture, etc.

La formulation générale de la fonction du poids que nous proposons est la suivante :

$$w(u, v) = exp\left(\frac{-\left\|f(u) - f(v)\right\|_2^2}{2\sigma_p^2}\right) . exp\left(\frac{-\left\|\mathcal{F}(f, u), \mathcal{F}(f, v)\right\|_2^2}{\sigma_s^2}\right) \tag{2.24}$$

Cette mesure peut incorporer aussi bien des informations topologique que photométrique. Les paramètres σ_p et σ_s sont des paramètres fixés à priori ou estimés localement par l'écart type mesuré sur la fonction initiale f^0. σ_p contrôle la similarité ou la proximité (cas $f(v) = v$) spatio-temporelle entre les noeuds et σ_s contrôle la similarité entre les patchs autour des noeuds.

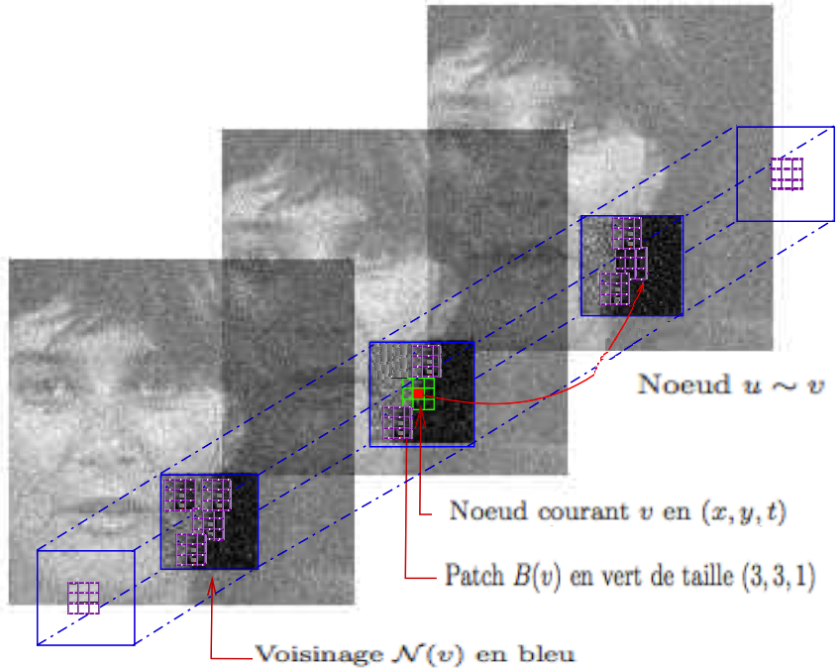


FIGURE 2.2: Exemples de voisinage spatio-temporel non local. Le pixel central (en rouge) est caractérisé par un vecteur de niveaux de gris défini grâce à un patch de taille $3 \times 3 \times 1$.

Dans l'approche locale, on considère que $F(f, v) = f(v)$. Le poids est une simple mesure de la différence entre les valeurs de $f(u)$ et $f(v)$. La figure 2.2 montre un exemple de patchs pour des images consécutives. Le pixel central (rouge) est caractérisé par un vecteur défini par un patch de la séquence d'images. Par souci de simplicité, nous considérons un patch de taille $(3 \times 3 \times 1)$. Le patch centré sur le noeud courant est affiché en vert. Les patchs voisins sont représentés en violet. Le voisinage spatio-temporel est affiché en bleu. Avec cette nouvelle représentation, l'espace des patchs est alors un espace de grande dimension ($\mathbb{R}^{r_x \times r_y \times r_t}$).

La figure 2.3 illustre la discrimination des patchs dans la fenêtre avec une approche non locale.

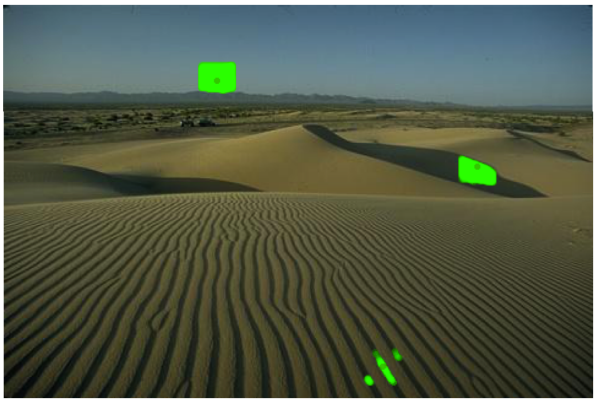


FIGURE 2.3: Pertinence de l'approche non locale. Les poids (en vert) représentent les patchs ayant un poids important ($\geqslant 0,5$) avec le patch central (rond). Patch : 5×5 et fenêtre 31×31.

Liens avec d'autres méthodes :

De manière non exhaustive, en considérant des valeurs particulières des paramètres, le processus de diffusion discrète 2.21 permet de retrouver plusieurs filtres et algorithmes de restauration utilisés en traitement de séquences d'images.

⋆ Quand le paramètre $\lambda \neq 0$ et que la fonction de poids w est constante, alors l'equation 2.21 correspond à la version discrète des équations aux dérivées partielles dans le domaine continu.
  - Quand $p = 1$, on retrouve la régularisation basée sur la variation totale(VT) [ROF92a] [GO08].
  - Quand $p = 2$, le système correspond aux fonctionnelles non locales proposées par Gilboa et Osher [GO07b].

⋆ Quand il n'y a pas d'attache aux données ($\lambda = 0$) et $p = 2$, nous retrouvons des filtres bien connus de la littérature :
  - Le filtre à moyennes non locales [BCM08] [SMC08] :

$$w(u,v) = exp\left(\frac{-\left\|\mathcal{F}(f,u) - \mathcal{F}(f,v)\right\|_2^2}{\sigma^2}\right) \tag{2.25}$$

  - Le filtre bilatéral [TM98] :

$$w(u,v) = exp\left(\frac{-\left\|u - v\right\|_2^2}{2\sigma^2}\right) . \; exp\left(\frac{-\left\|f(u) - f(v)\right\|_2^2}{\sigma^2}\right) \tag{2.26}$$

2.2.5 Résultats et performances

L'application de l'algorithme 2.21 présente un coût de calcul élevé puisqu'il s'agit de comparer le patch centré sur chaque sommet v avec tous les patchs contenus dans le voisinage non local $\mathcal{N}(v)$. Les séquences d'images sont dégradées par un bruit blanc gaussien additif η de variance σ^2. Nous avons mené une série d'expérimentations et nous avons étudié les points suivants :

Optimisation : Pour optimiser le temps de calcul, nous avons choisi d'utiliser 30% de voisins hors du patch parmi $\mathcal{N}(v)$. Les paramètres utilisés : Patch $3 \times 3 \times 3$, une taille de fenêtre de $7 \times 7 \times 3$ $p = 2$ et $\lambda = 0.5$. La série de tests sur des séquences corrompues par des bruits différents (voir figure 2.4) montre comment le PSNR diminue considérablement lorsque le niveau de bruit augmente. Ces mesures confirment également que la méthode non locale optimisée reste un bon compromis qui donne des résultats satisfaisants.

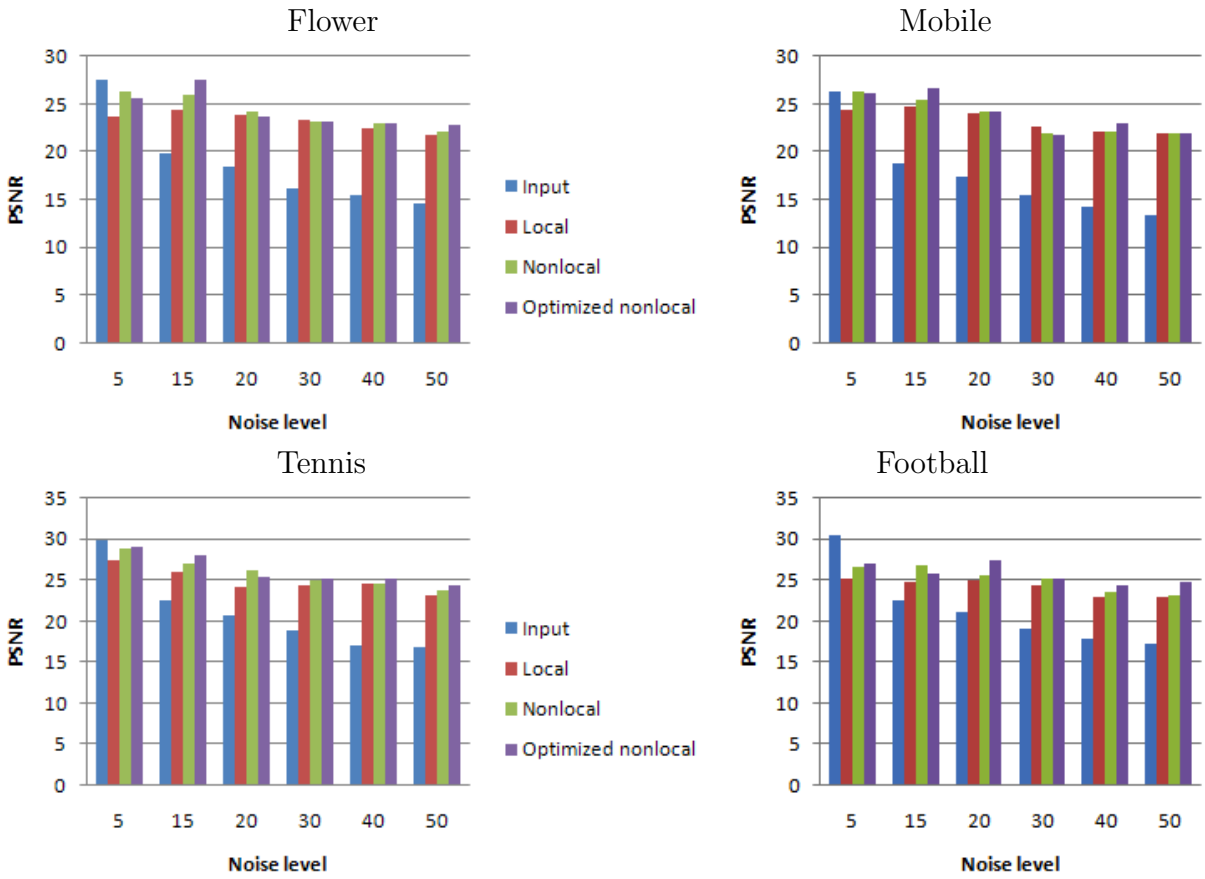


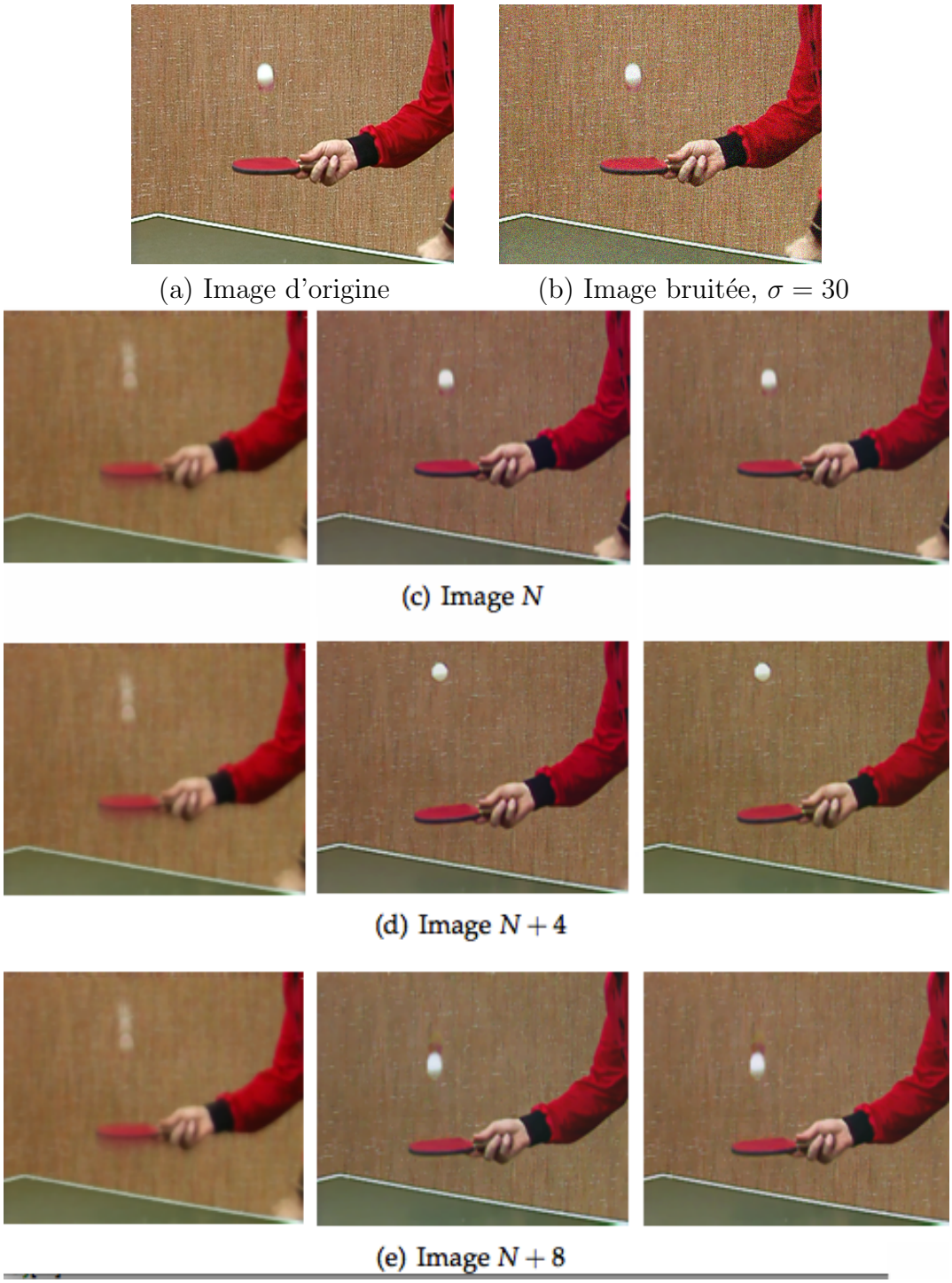
FIGURE 2.4: Comparaison de PSNR des approches de débruitage (locales, non locales, non locales optimisées) et le PSNR d'entrée.

Débruitage 2D/3D : Nous avons d'abord comparé l'approche de débruitage 2D (image par image) et 3D (séquence entière). Le PSNR des résultats 3D est systématiquement plus élevé que le PSNR des résultats 2D ce qui confirme l'intérêt de la redondance temporelle pour le débruitage vidéo.

Débruitage Local/Non Local : La figure 2.5 montre que la méthode non locale en 2d+t fournit de meilleurs résultats visuels que l'approche locale. Les observations visuelles sont également confirmées par les mesures PSNR rapportées dans le graphique 2.4.

Influence de p : Pour mettre en évidence l'influence de p sur les résultats de débruitage, nous avons utilisé une fonction de poids constante $w = 1$. Quand le paramètre de régularité $p = 2$, le filtrage lisse l'ensemble des données y compris dans les zones de fortes courbures. Nous avons constaté que quand $p = 2$ $(\forall\lambda)$ et $\lambda = 0$ $(\forall p)$, nous avons approximativement les mêmes effets de filtrage. Le traitement a tendance à moyenner les données.

Quand le paramètre $p = 1$, les courbures initiales sont mieux préservées. Le cas $p = 1$ agit comme une médiane. Dans la série d'expérimentations réalisées sur les séquences vidéos, nous avons pu observé qu'on obtient généralement un meilleur PSNR quand les valeurs des paramètres p et λ sont faibles.



(a) Image d'origine (b) Image bruitée, $\sigma = 30$

(c) Image N

(d) Image $N + 4$

(e) Image $N + 8$

FIGURE 2.5: Illustration de nos trois méthodes de débruitage avec $p = 2$, $\lambda = 0.5$ et $\sigma_s = 30$ après 10 itérations. De gauche à droite : le résultat local, le non local et le non local optimisé.

De manière générale, quand $p \leq 1$ ($p = 0,5$) des regroupements de noeuds ont tendance à devenir constants par morceaux. La figure 2.6 montre l'effet de cette simplification. Ces résultats peuvent être mis à profit pour la segmentation vidéo et la détection d'objets visuels.

(a) Original

(b) Bruitée, $\sigma = 10$

(c) Image N

(d) Image $N + 4$

(e) Image $N + 8$

FIGURE 2.6: Simplification de vidéos par la méthode non locale optimisée, avec $\lambda = 0$ et $w = 1$ après 5 itérations. De gauche à droite : p = 2, p = 0, 5 et p = 0, 1.

Comparaison avec trois méthodes non locales de la littérature : La première est celle de l'approche originale du filtre à moyennes non locales (NL-means) [BCM08]. La seconde est basée sur l'algorithme VBM3D [DFE07] qui utilise également une multitude de patchs dans le voisinage 3D de chaque pixel pour atténuer le bruit. Une transformée par ondelettes 3D est appliquée avec un filtrage de Wiener pour améliorer les résultats de débruitage. La troisième méthode proposée par Protter el al. [PE07] est basée sur la diffusion de dictionnaires parcimonieux. Les résultats sont présentés dans les tableaux 2.1 et 2.2. Avec la moyenne des PSNR du tableau 2.2, nous obtenons la deuxième performance moyenne de l'ordre de (29.02 dB) après l'approche basée sur l'apprentissage de dictionnaires parcimonieux (29,2 dB), suivi de la méthode des moyennes non-locales (28,86 dB) et la méthode VBM3D (27.91dB). Nous avons fixé les paramètres $\lambda = 1$ et $p = 1$, et nous nous sommes arrêtés après convergence à 10^{-3} près entre deux itérations consécutives.

σ	**Football** (180 × 144 × 105)		**Tennis** (216 × 172 × 126)		**Flower** (180 × 144 × 126)		**Mobile** (180 × 144 × 251)	
5	36.96	35.95	38.32	37.12	36.84	36.03	36.81	36.22
	37.42	36.95	38.16	**38.2**	36.73	**36.82**	**38.55**	38.41
15	30.74	29.61	32.46	29.56	30.16	29.04	30.27	29.43
	30.96	30.69	32.10	**32.15**	29.69	**30.17**	**31.62**	**31.62**
20	29.22	28.14	**30.76**	28.14	**28.58**	27.56	28.58	27.73
	29.39	29.20	30.51	**30.76**	28.03	28.17	**29.78**	29.11
30	27.19	26.52	27.86	26.68	**26.26**	25.33	25.98	25.36
	27.23	27.2	**28.43**	27.93	25.56	26.24	**26.87**	26.0
40	25.84	24.97	26.45	25.93	24.49	23.51	24.04	23.60
	25.86	25.85	**27.22**	26.41	24.02	**24.55**	**25.11**	25.05
50	**24.81**	24.11	25.79	25.34	22.48	22.16	22.71	21.93
	24.73	24.73	26.34	**26.35**	**22.80**	22.6	**23.64**	23.69

TABLE 2.1: Comparaison du PSNR (dB), des séquences bruitées (avec différents σ). De haut en bas et de gauche à droite : NL-Means, VBM3D, la diffusion de dictionnaire, et notre méthode(avec $p = 2$, $\lambda = 50$ and $\sigma_s = 30$). Le meilleur résultat est mis en caractères gras.

σ	**5**		**15**		**20**	
PSNR Moyen	37.23	36.33	30.91	29.41	29.29	27.89
	37.72	37,6	31.09	**31,16**	**29.43**	29,31
σ	**30**		**40**		**50**	
PSNR Moyen	26.82	25.97	25.20	24.50	23.70	23.38
	27.02	26,84	**25.55**	25,46	**24.38**	24.34

TABLE 2.2: PSNR moyen par σ pour chacune des 4 approches.

2.2.6 Conclusion

Dans cette section, nous avons étudié la problématique du débruitage de séquences d'images. Nous avons présenté notre contribution concernant le débruitage vidéo par une approche variationnelle basée sur une régularisation non locale sur graphe par p-Laplacien avec $p \in]0, +\infty[$. Cette approche présente un avantage par rapport aux méthodes classiques puisque la formulation de la régularisation est exprimée sur graphe et en discret. Elle permet donc d'adapter les traitements en modifiant simplement le voisinage des noeuds et la connexité du graphe. Nous avons pu montrer les liens de cette approche avec différentes méthodes de la littérature. Les paramètres de régularisation sont choisis a priori. Pour augmenter la rapidité de traitement, nous avons fait le choix de n'utiliser qu'un pourcentage (fixe) de noeuds dans le voisinage. La fonction du poids w capte les interactions locales et non locales entre les patchs comparés et permet d'obtenir de bons résultats. Le choix de la méthode non locale s'impose face à la méthode locale. Les paramètres σ_k et les dimensions de la fenêtre de recherche et du patch sont dépendants du contenu à traiter. Les variantes multi-résolutions permettent une certaine automatisation mais le réglage manuel d'un expert reste la meilleure solution pour ne pas alourdir le processus. Nous n'avons pas étudié l'influence de p sur le débruitage. Mais, nous avons remarqué que le débruitage par $p \leq 1$ agit comme un processus de simplification qui peut être utilisé comme un pré-traitement important facilitant la segmentation.

Dans la section suivante, nous nous intéressons à un deuxième type de restauration. Il s'agit de l'inpainting qui consiste en l'interpolation de données manquantes.

2.3 Inpainting

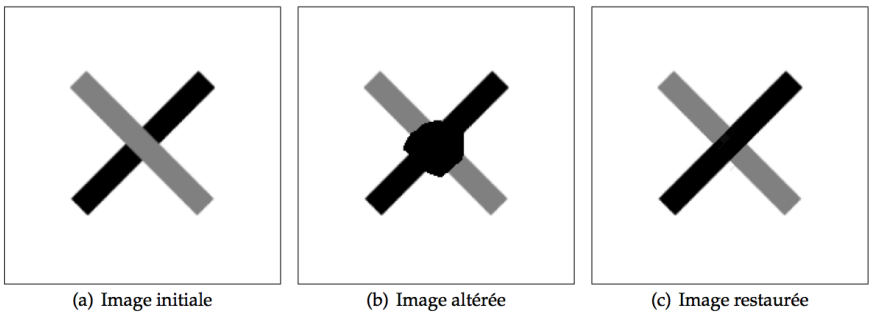
☞ Publications associées : [RI6, CI5, PS1] [ThèseF1] [GEL11]

Dans cette section, nous nous intéressons à la restauration des données manquantes. Contrairement à la restauration des données bruitées, il n'y a pas de bruit à modéliser ni de données initiales auxquelles il faudrait être fidèle. Il s'agit ici d'un problème d'interpolation qui consiste à réparer une scène en construisant un contenu nouveau en harmonie avec les données non perdues de la scène en question. Nous n'abordons pas l'étape préliminaire de sélection des défauts ou des dégradations telles que les déchirures ou les taches. Nous considérons uniquement le remplissage du trou créé par la suppression des taches et supposons que la région à combler est déjà définie. Nous présentons nos deux contributions concernant l'inpainting. Dans un premier temps, à la section 2.3.2, nous présentons l'extension discrète de la méthode d'interpolation continue de Chan et Shen[CS02] au domaine de l'inpainting. Nous traitons le résultat d'une première interpolation par régularisation itérative afin d'améliorer la qualité de l'interpolation. À la section 2.3.3, nous nous intéressons aux fonctions p-harmoniques qui constituent un cadre général d'interpolation. Ce dernier permet la réalisation de nombreuses applications dont l'inpainting sur lequel nous allons nous focaliser avec le cas $p = 2$.

2.3.1 Introduction et problématique

De nombreuses applications se basent sur l'inpainting. À titre non exhaustif, nous pouvons citer la restauration de photographies, l'agrandissement et l'interpolation d'images, la compression d'images, la correction des zones perdues en cas d'erreur de transmission, la suppression d'objets ou de textes pour des effets spéciaux, etc.

L'inpainting consiste à remplir les zones manquantes d'une image ou d'une vidéo avec un contenu approprié calculé à partir des données préservées. Il est important de noter que l'objectif est de construire un contenu plausible et qu'il peut y avoir plusieurs résultats différents acceptables. La figure 2.7 illustre un cas où le résultat de l'inpainting est vraisemblable mais diffère de l'image initiale. Cet exemple souligne l'influence que peut avoir la construction du masque.



(a) Image initiale (b) Image altérée (c) Image restaurée

FIGURE 2.7: Influence du masque et Inpainting : Après l'inpainting, la bande noire passe au-dessus de la bande grise (c) alors qu'à l'origine c'est l'inverse.

L'inpainting a fait l'objet d'un grand nombre de travaux. Différentes stratégies ont été envisagées pour proposer une solution à ce problème. On distingue trois grandes classes de méthodes :

1) La première s'est focalisée sur la reconstruction de **structures géométriques**. Parmi les méthodes proposées, un certain nombre sont variationnelles ou basées sur les EDP. Les problèmes d'optimisation sont résolus avec la contrainte des données connues sur le bord du trou. Différents modèles d'images sont utilisés tels que le modèle VT [ROF92b], [CS05b], le modèle Mumford Shah [MS89], ou une fonctionnelle elastica d'Euler [MM98]. Il existe également des méthodes dans lesquelles une EDP est directement définie comme dans les travaux de Ballester et al. [BSCB00a] et de Bertalmio et al. [BBS01]. Dans le même esprit, Chan et Shen [CS02] proposent un modèle d'inpainting basé sur la VT et fondé sur les principes Bayesiens et variationnels :

$$\min_f \quad E_{TV}(f) = \int_\Omega |\nabla f| \tag{2.27}$$

avec la contrainte :

$$f = g \text{ sur } \partial A \tag{2.28}$$

où A est la zone à combler et ∂A sa frontière.

Minimiser cette fonctionnelle d'énergie revient à relier les lignes de niveau de part et d'autre de la région à combler en suivant la plus petite distance. Ceci peut être exprimé par la formule co-aire :

$$\int |\nabla f| dx = \int_0^1 \int_{\Gamma_\lambda} ds d\lambda \tag{2.29}$$

où $\Gamma_\lambda = \{u : f(u) = \lambda\}$ est la ligne de niveau et ds est la longueur de l'arc des lignes de niveau. La variation totale de u est la longueur totale de toutes les lignes de niveau. Par conséquent, avec la contrainte (2.28), les contours nets sont reliés suivant les lignes de niveau dans la région connue. L'inpainting par VT interpole des images dans toutes les régions manquantes tout en préservant les contours nets. Toutefois, il ne relie pas toujours les contours correctement.

Chan et al. [CKKS02] [CS05b] ont proposé de minimiser l'énergie de l'elastica d'Euler au lieu de la VT. L'inpainting par l'elastica d'Euler améliore l'inpainting VT parce que la courbure des lignes de niveau est davantage pénalisée. Ceci permet d'éviter les entortillements aux bords du trou en raison de leur courbure infinie. L'inpainting par l'elastica d'Euler répare les zones manquantes selon la minimisation d'énergie suivante :

$$\min_f \quad E_{elastica(f)} = \int (a + b\kappa^2)|\nabla f| \quad \text{avec la contrainte f=g sur } \partial A \tag{2.30}$$

Dans cette fonctionnelle, a et b sont des constantes positives et $\kappa = \nabla.(\frac{\nabla f}{|\nabla f|})$ est la courbure de f. Minimiser cette fonctionnelle d'énergie revient à relier des contours nets en fonction de la courbure des lignes de niveau dans la région connue. Ceci peut être expliqué par la formule de la co-aire :

$$\int (a + b\kappa^2)|\nabla f| = \int_0^1 \int_{\Gamma_\lambda} (a + b\kappa^2) ds d\lambda \tag{2.31}$$

Il existe d'autres méthodes qui impliquent des EDP rigides du quatrième ordre dans le problème d'inpainting. Burger et al. [BHS09] ont proposé un modèle pour l'inpainting des images en modifiant l'équation de Cahn-Hilliard. Le modèle à deux échelles de Cahn-Hilliard réussit à relier les contours à travers de larges régions. Récemment, une catégorie de modèles basés sur les ondelettes ont été proposés par Dobrosotskaya et Bertozi [DB08].

Généralement, ces méthodes offrent les meilleurs résultats sur les régions de petites tailles. Inversement, lorsque les alentours du trou sont formés de structures stochastiques comme une texture, ce type de méthodes tend à fournir un contenu dégradé. Ce défaut est dû à la nature locale de ces méthodes, inadaptées à la reconstruction de structures stochastiques.

2) Le deuxième groupe d'algorithmes d'inpainting est basé sur la reconstruction des régions manquantes par **synthèse de textures** grâce à des modèles statistiques du contenu de l'image. L'échantillonnage non-paramétrique proposé par Efros et Leung [EL99] est utilisé dans les techniques basées textures. Les méthodes de rééchantillonnage non-paramétrique, dites également méthodes par l'exemple [ZWM97], [CPT03] [TD05] ont prouvé leur capacité à reconstruire des structures stochastiques dans les zones occultées. Ces méthodes sont basées sur la recherche et la copie de patchs similaires dans le voisinage du trou. Les données se reconstruisent du bord extérieur au centre du trou. En général, une stratégie de remplissage est employée. Celle-ci influence le résultat obtenu. L'approche proposée par Criminisi et al. [CPT03] se base sur une mesure de gradient au bord du trou afin de déterminer l'ordre de remplissage. Les patchs des zones à fort gradient sont traités en premier.

Les méthodes par l'exemple, par le caractère local de leurs définitions, sont limitées dans la restauration de la géométrie à grande échelle. Des approches multi-échelles peuvent minimiser cet inconvénient. L'inconvénient de cette démarche est de ne pas pouvoir rectifier le remplissage dans le cas où un patch non optimal aurait été choisi.

3) La troisième classe est basée sur les **méthodes non locales** qui tentent d'unifier les deux approches précédentes [PBC08]. Il est important de rappeler que les images naturelles comportent aussi bien des structures géométriques que des structures stochastiques. Les travaux récents sur la non localité ont mis en évidence que la multitude des candidats utilisés pour le calcul de la valeur d'un nœud est un avantage à exploiter pour obtenir des reconstructions plus rigoureuses. Depuis les travaux sur le filtrage NL-means, de nombreuses méthodes variationnelles et basées sur les EDP ont été proposées. Elles ont révélé leur efficacité à préserver les structures stochastiques. Kindermann et al. [KOJ05] ont été les premiers à interpréter le NL-means comme une régularisation basée sur des fonctions non locales. Gilboa et Osher [GO07b] ont proposé une approche de régularisation non locale continue avec des applications au filtrage et à l'inpainting. Un cadre de travail variationnel d'inpainting non local a été aussi présenté dans [ACS09].

2.3.2 Inpainting par régularisation

Dans cette section, nous présentons l'extension de notre cadre de régularisation discrète à l'inpainting des images et des vidéos. Cela correspond à l'extension de la méthode continue de [CS05b] au domaine discret.

2.3.2.1 Formulation variationnelle

L'inpainting peut être considérée comme une spécialisation du cadre de régularisation présentée à la section 2.2. Au démarrage du processus d'inpainting, de nouvelles valeurs sont calculées pour les parties manquantes. Le paramètre de fidélité λ est fixé à 0.

Soit $A \subset V$ l'ensemble des noeuds correspondant aux trous à remplir (voir figure 2.8). Dans le cas de l'inpainting, on ne cherche à restaurer que les données manquantes (noeuds appartenant à A)

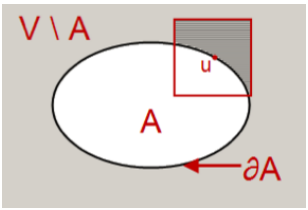


FIGURE 2.8: Illustration de l'inpainting de la zone A à partir des données connues (V\A) en comparant le patch centré autour de u (la zone hachurée) avec les patchs de même forme.

Rappelons la formulation de régularisation discrète générale (2.18) qui est :

$$f^* = \underset{f \in \mathcal{H}(V)}{min} \left\{ \frac{1}{p} \sum_{v \in V} \left\| \nabla f(v) \right\|^p + \frac{1}{2} \sum_{v \in V} \lambda(v) \left\| f^0(v) - f(v) \right\|^2 \right\} \text{ avec } p \in]0, +\infty[$$

avec :

$$\lambda(v) = \begin{cases} \text{constante} & \text{si } v \in V \backslash A \\ 0 & sinon \end{cases} \tag{2.32}$$

Pour l'estimation des premières valeurs, cette formulation se réduit à :

$$f^* = \underset{f \in \mathcal{H}(V)}{min} \left\{ \frac{1}{p} \sum_{v \in V} \left\| \nabla f(v) \right\|^p \right\} \tag{2.33}$$

Lorsque chaque élément de A ait été évalué une première fois, le processus de régularisation est itéré avec $\lambda \geqslant 0$ afin de raffiner les valeurs obtenues par l'interpolation.

Liens avec d'autres méthodes :

Le cadre ci-dessus unifie et englobe plusieurs techniques spécifiques présentes dans la littérature. En fait, en considérant des valeurs particulières des paramètres, nous retrouvons les résultats qui ont été établis dans le traitement de l'image.

⋆ Pour $p = 2$ et une itération, la méthode non locale est équivalente au filtre à moyennes non locales [BCM08] qui a été adapté à l'inpainting par [WO08].

⋆ Avec $p = 1$ et $w = 1$, on obtient l'inpainting par variation totale anisotrope. Ce modèle a été appliqué en inpainting par [CS05b].

⋆ Notre méthode peut être considérée comme une extension de la méthode de [EL99]. En effet, si nous construisons le graphe des $k - ppv$ avec $k = 1$ et une distance de patchs entre les nœuds, nous obtenons la même approche. Toutefois, dans notre algorithme, nous pouvons considérer des valeurs différentes pour k.

2.3.2.2 Résultats et Discussion

La méthode consiste à remplir un trou en partant de son contour extérieur jusqu'à son centre de manière récursive, en utilisant une carte de distance discrète. L'approche de remplissage par contours imbriqués est un choix judicieux par rapport à un balayage linéaire car ce dernier favorise les effets de bord et la convergence est beaucoup plus lente.

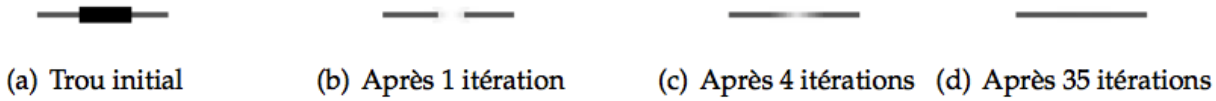
Pour un noeud donné (pixel ou voxel) du contour extérieur, nous calculons une nouvelle valeur à partir des données connues dans une fenêtre de recherche centrée sur le sommet courant en évitant les patchs "vides". Nous ne mettons à jour l'ensemble A que lorsqu'un contour a été complètement traité. Nous n'incluons pas la nouvelle valeur calculée d'un sommet dans l'estimation des autres sommets appartenant au même contour. Ceci permet de réduire le risque de propagation d'erreurs dans les calculs. La distance entre le patch de référence et les autres dans la fenêtre de recherche tient compte de la forme du patch traité et ne considère que les patchs de même forme.

La stratégie adoptée consiste à :

Etape 1 : Itérer la régularisation de chaque contour imbriqué dans le trou avant de commencer le remplissage du contour imbriqué suivant.

Etape 2 : Utiliser une stratégie multi-résolution pour le choix du patch et améliorer la discrimination des patchs candidats. En fait, on garde le patch qui a le maximum de votes comme patch candidat (cf fig 2.12).

Le remplissage itératif par régularisation du trou présente le grand avantage de pouvoir corriger un premier choix inapproprié ou d'améliorer un choix perfectible comme l'illustre la figure 2.9 de reconstruction d'un trait.



(a) Trou initial (b) Après 1 itération (c) Après 4 itérations (d) Après 35 itérations

FIGURE 2.9: Correction progressive

Influence de σ_s : C'est un paramètre d'échelle comme dans toutes les méthodes à noyau. Si $sigma_s$ est trop grand, le résultat est flou. Il faut donc l'ajuster comme l'on règle la focale d'une lentille afin d'obtenir une image nette. Différentes heuristiques ont été proposées et la question reste ouverte.

En utilisant une méthode de restauration non locale par patchs, nous arrivons à restituer aussi bien les contours que la texture. Une sélection de plusieurs exemples montrent l'efficacité de notre approche d'inpainting par régularisation. sur des images fixes et sur des séquences vidéo.

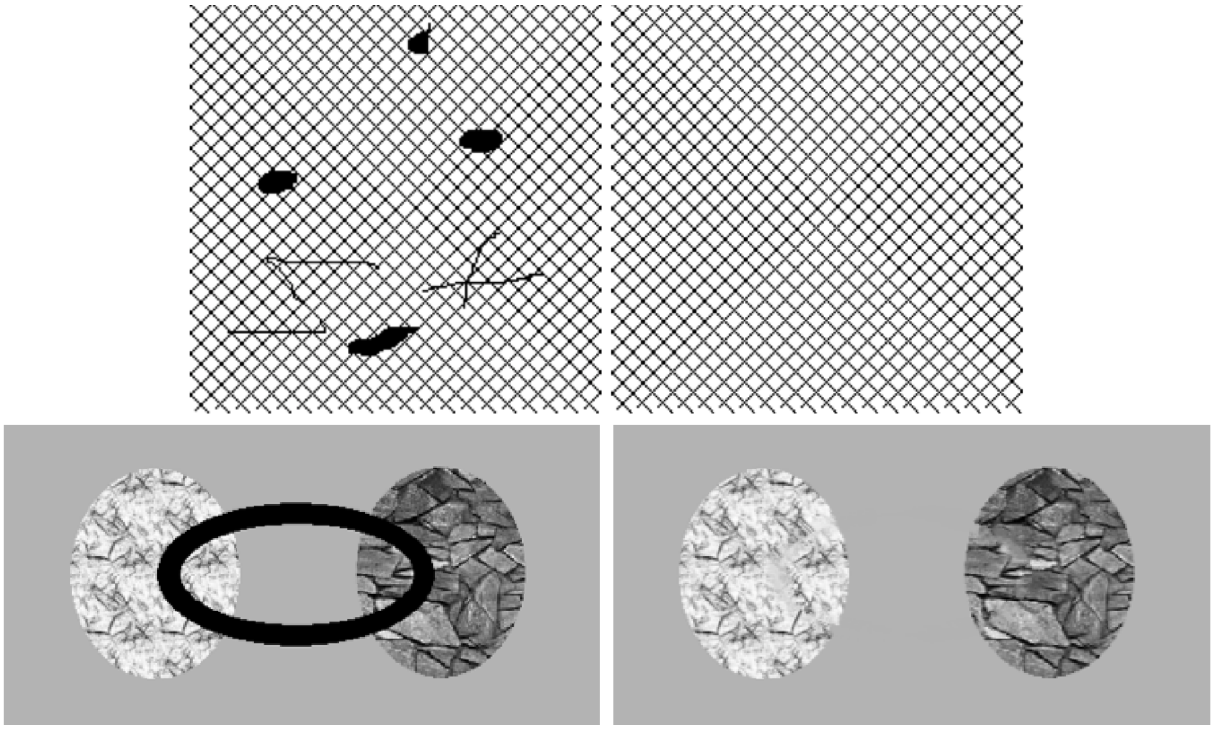
Généralement, de larges patchs sont adéquats pour les zones homogènes, tandis que les zones texturées ont tendance à exiger un patch de taille variable (petite) en fonction du contenu représenté par la nature des motifs.

FIGURE 2.10: Illustration de patchs candidats. Les points (en bleu) ayant un poids important (supérieur à 0.5) avec le point traité (en rouge) respectent la forme des contours de la zone traitée et des zones manquantes. Paramètres : Fenêtre 61 × 61 et patch 13 × 13.

Inpainting d'images :

Pour les images en couleur, les composantes RGB sont traitées indépendamment. Parfois, certaines aberrations chromatiques peuvent apparaître. Une solution qui a été adoptée consiste à lier les trois composantes par un modèle couleur [BSCB00b].

FIGURE 2.11: Restauration d'une image couleur réelle avec $p = 2$, une fenêtre de recherche 21×21 et un patch 11 ×11



(a) Image altérée (b) Image restaurée

FIGURE 2.12: Restauration avec $p = 2$, une fenêtre de 15×15 et un patch 3×3.

La figure 2.12 illustre l'efficacité de notre algorithme concernant la reconstruction de la forme géométrique et de la texture synthétique. D'autres résultats sur des images de textures sont rapportés sur les figures 2.14 et 2.13.

(a) Image altérée (b) Image restaurée

FIGURE 2.13: Inpainting avec $p = 2$, une fenêtre de recherche de 17 ×17 et un patch 7 × 7.

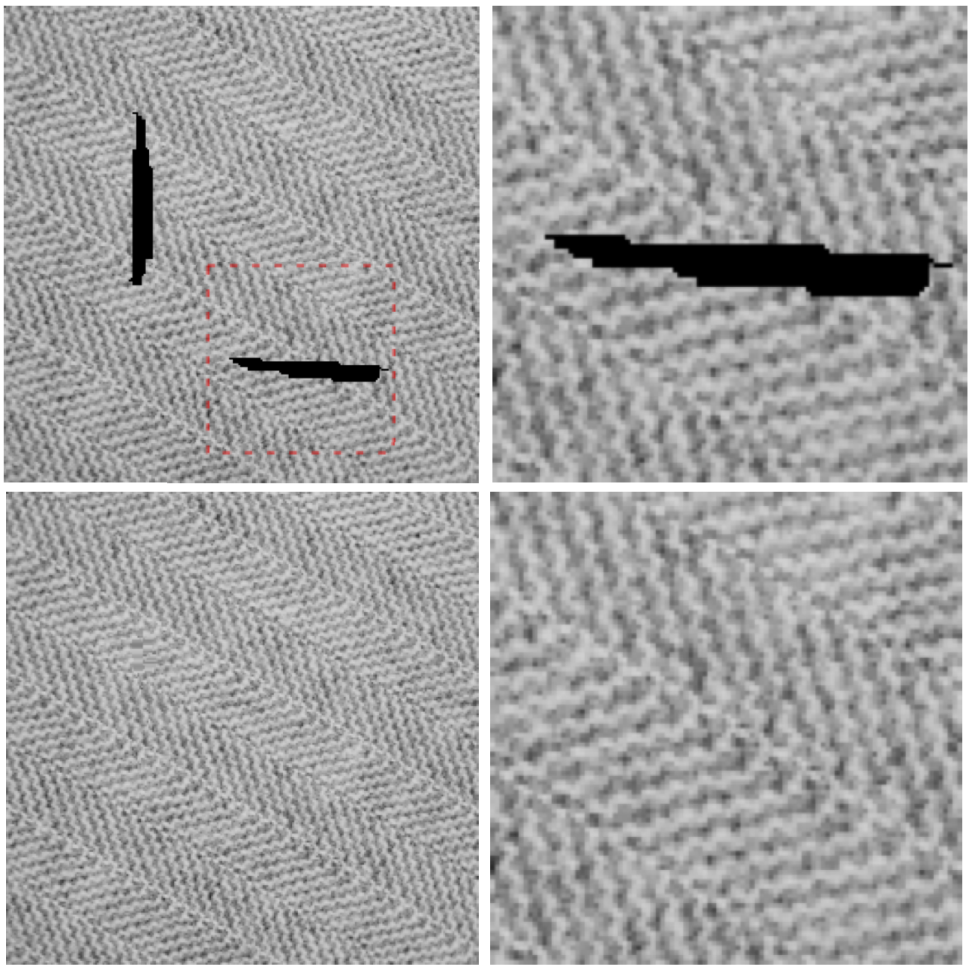


FIGURE 2.14: Inpainting d'une image texturée (weave) avec une fenêtre de 61×61 et un patch 31×31. De gauche à droite : sur la première ligne, l'image corrompue et une zone agrandie de l'image. Sur la seconde ligne, le résultat de l'inpainting non local avec $p = 2$ et un zoom sur la région.

Inpainting vidéo

Il est possible d'observer la contribution de la redondance temporelle dans le processus d'inpainting. En effet, dans la figure 2.15, nous avons altéré quelques frames de la vidéo de Suzie (une frame sur trois). Le zoom sur les régions restaurées montre la grande efficacité de notre approche non locale à reconstruire la texture et les structures géométriques. L'approche peut aussi être très utile pour supprimer des objets, comme l'illustrent le figures 2.17 et 2.16. Nos tests sur plusieurs corpus vidéo montrent que la suppression de plusieurs petites zones est plus facile que la suppression d'une grande zone. En effet, dans la mesure où les petites régions sont redondantes dans la vidéo, de nombreuses parties connues existent tout autour et pourraient être utilisées pour reconstituer les parties manquantes. Au contraire, quand une grande zone compacte manque, la plupart des candidats utilisés pour remplir la partie intérieure du trou sont des valeurs reconstruites, ce qui conduit à une moindre qualité. Une façon de surmonter cette difficulté est de considérer une fenêtre de recherche importante, mais qui a un coût de calcul plus élevé.

FIGURE 2.15: Sur la 1ère ligne, la frame altérée et le résultat de l'inpainting non local avec une fenêtre 21×21 et un patch 6×6. La dernière ligne présente un zoom sur des régions restaurées.

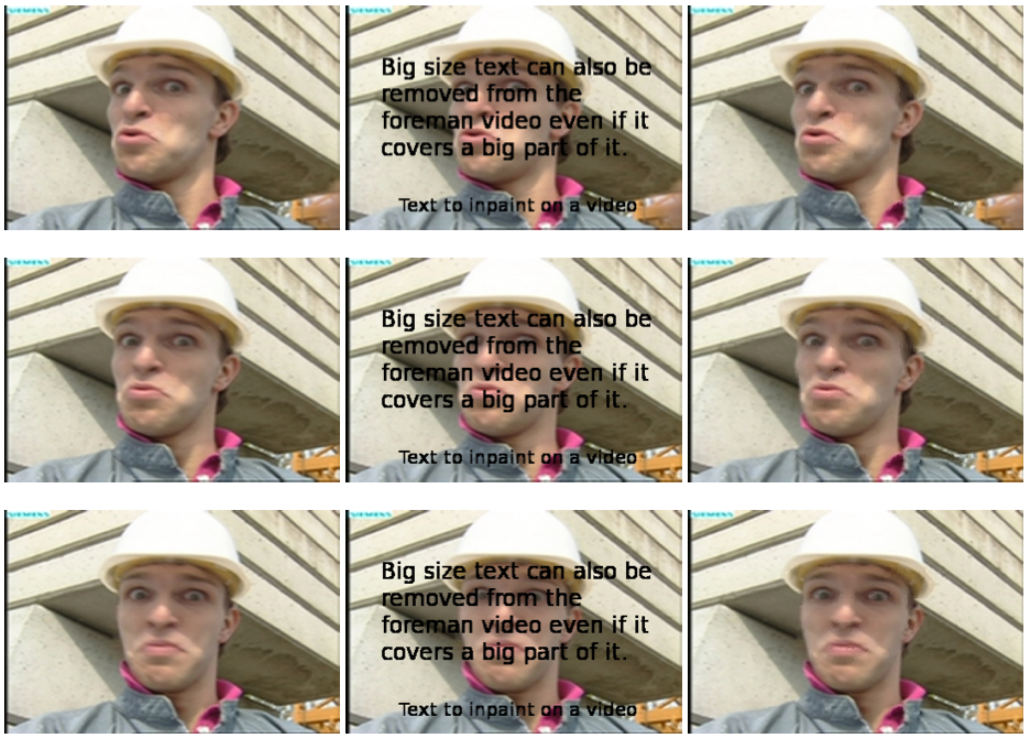


FIGURE 2.16: Suppression de texte d'une séquence de la vidéo foreman. De gauche à droite, la séquence d'origine, la séquence corrompue et le résultat de restauration non local.

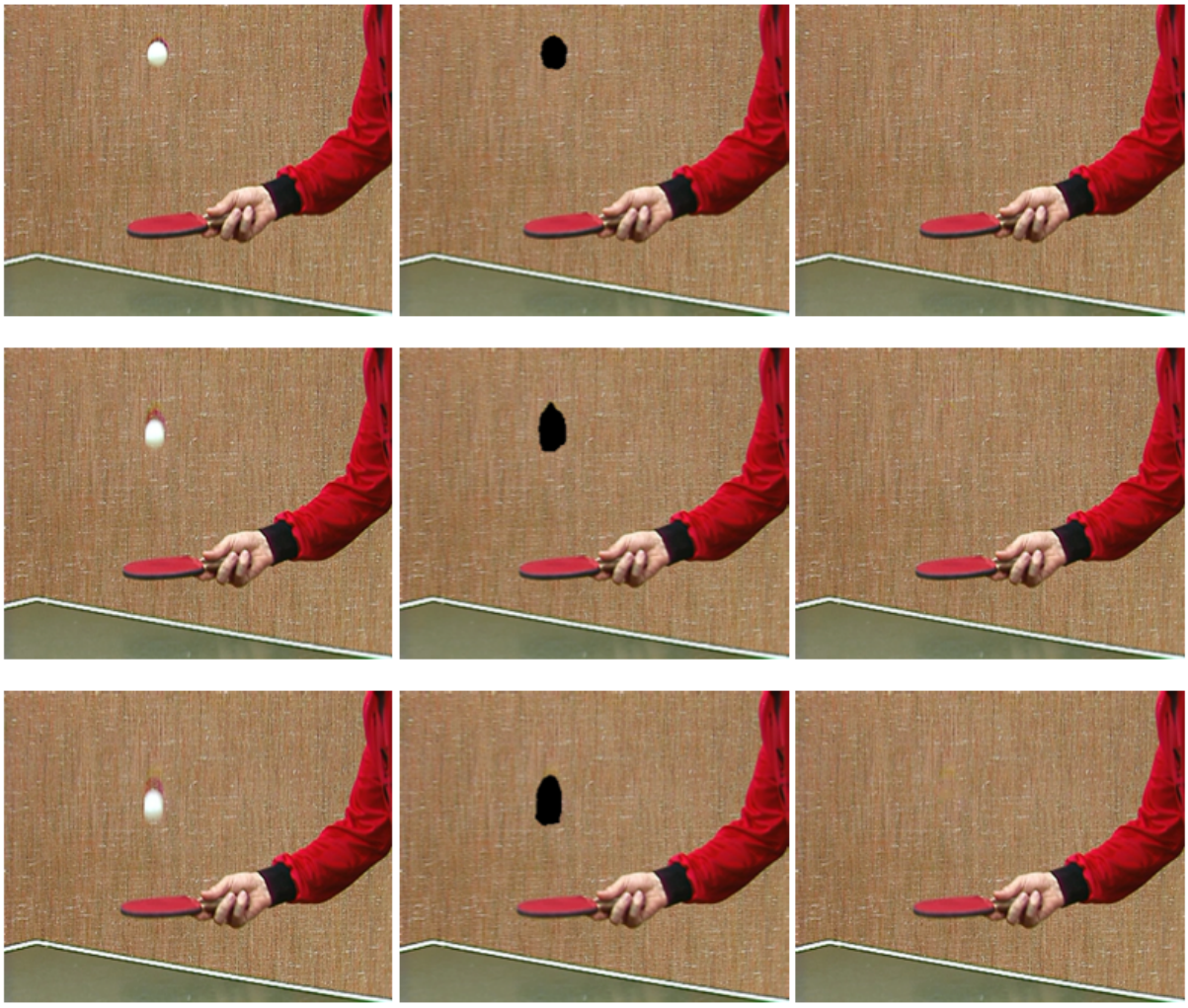


FIGURE 2.17: Suppression d'objets dans la séquence vidéo tennis (de table). De gauche à droite, une séquence d'origine, la séquence corrompue et le résultat de restauration non local.

2.3.3 Interpolation sur graphe

À la section précédente, nous avons présenté l'extension des méthodes d'interpolation continues au domaine discret. Ici, nous nous intéressons aux fonctions harmoniques sur graphe.

2.3.3.1 Fonctions p-harmoniques

Je présente ici une nouvelle manière de résoudre le problème p-Dirichlet avec des fonctions p-harmoniques sur les graphes de topologie arbitraire. Mathématiquement, les problèmes d'interpolation consistent à trouver une fonction p-harmonique f qui minimise la p-énergie avec des conditions de bord (des valeurs spécifiques sur la frontière ∂A).

Définition : Soit $A \subset V$. Une fonction f est p-harmonique sur A si elle minimise

l'énergie :

$$\mathcal{R}_w^p(f, A) = \frac{1}{2p} \sum_{\substack{u \sim v \\ v \in A}} \left\| \nabla f(v) \right\|_p^p \tag{2.34}$$

pour toute fonction f dans $A \cup \partial A$, avec la contrainte de bord $f = g$ sur ∂A.

Une manière de résolution le problème 2.34 est de considérer l'équation suivante :

$$\left\{ \begin{array}{rcll} \Delta_{p,w}^a f(u) & = & 0 & \forall u \in A \\ f(u) & = & g(u) & \forall u \in \partial A \end{array} \right. \tag{2.35}$$

En définissant $\gamma(u, v) = w(u, v)^{p/2} \left| f(u) - f(v) \right|^{p-2}$, l'équation 2.35 peut être formulée :

$$\Delta_{p,w}^a f(u) = \sum_{v \sim u} \gamma(u, v) \left(f(u) - f(v) \right) = 0 \tag{2.36}$$

Nous obtenons alors comme point fixe :

$$f(u) = \frac{\sum\limits_{v \sim u} \gamma(u, v) f(v)}{\sum\limits_{v \sim u} \gamma(u, v)} \tag{2.37}$$

La méthode itérative de Gauss-Jacobi conduit à une méthode itérative robuste pour résoudre le problème de p-Dirichlet avec des fonctions p-harmoniques sur graphes comme celles que nous avons présentées dans la section 2.2.3. Un des principaux avantages de l'approche est l'unification des approches locales et non locales dans un cadre discret qui le rend applicable à toutes les données représentées sur graphes.

2.3.3.2 Résultats préliminaires

Ces travaux ont été effectués dans le cadre de la thèse de M. Ghoniem, les résultats présentés sont très encourageants comme peuvent le montrer les figures suivantes. La figure 2.18 fournit un exemple d'inpainting sur une image synthétique simple.

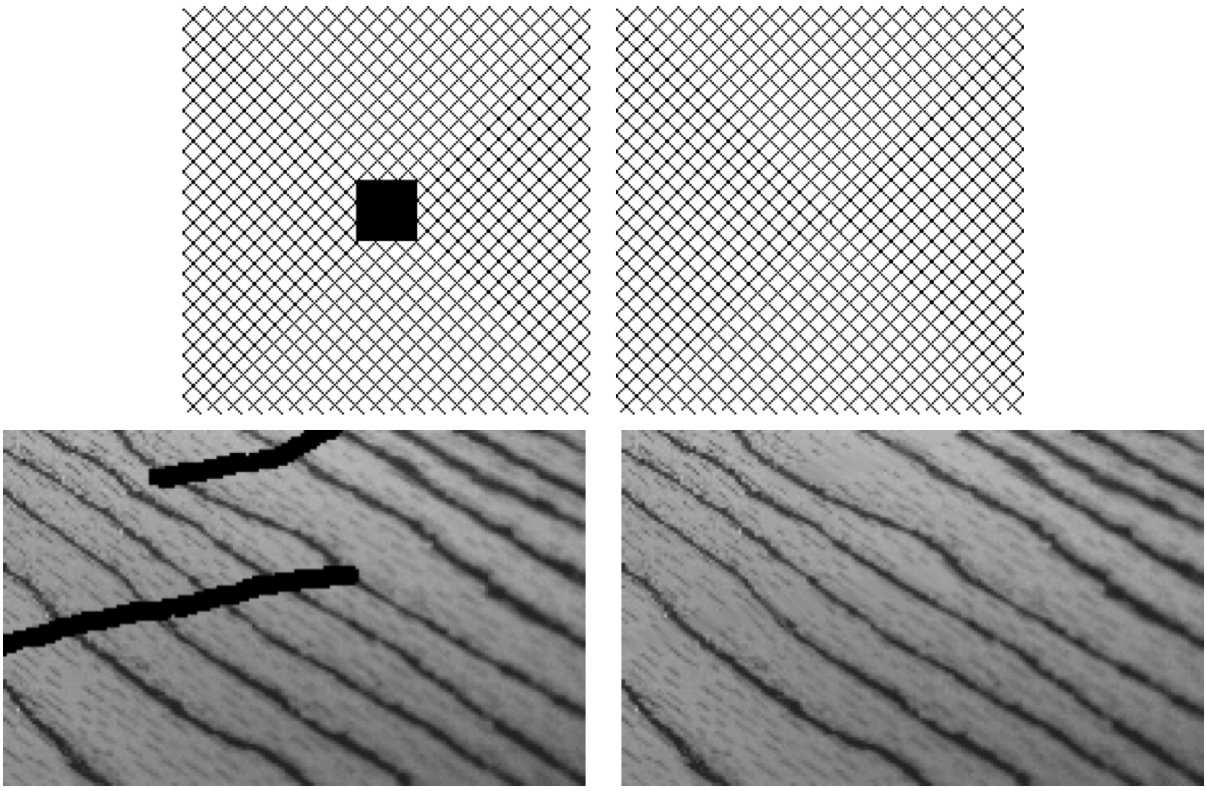


FIGURE 2.18: Restauration p-harmonique avec une fenêtre de recherche 3×31 et un patch 15×15. De gauche à droite : L'image corrompue et le résultat de l'inpainting local avec $p = 2$.

La figure 2.19 présente une comparaison visuelle entre les deux méthodes de restauration p-régularisation et p-harmonique pour le cas de $p = 2$.

2.3.4 Conclusion

Nous avons présenté nos contributions à la problématique de l'inpainting et l'interpolation des images et des vidéos. Nous avons présenté deux approches d'inpainting. La première est l'extension de l'interpolation continue au domaine discret. Nous traitons le résultat d'une première interpolation par régularisation itérative afin d'améliorer la qualité de l'interpolation. La seconde approche est basée sur les fonctions p-harmonique. De nombreuses applications en découlent : la suppression de textes et d'objets, ou au contraire l'incrustation d'objets.

Nous avons présenté un modèle variationnel discret avec une régularisation itérative et proposé une nouvelle approche pour résoudre le problème p-Dirichlet avec des fonctions p-harmoniques sur les graphes de topologie arbitraire qui ouvre de nouvelles perspectives en traitement non local des données en vidéo.

FIGURE 2.19: Image de Barbara avec $p = 2$, une fenêtre de recherche 21×21 et un patch 5×5. De gauche à droite : 1ère ligne : l'image d'origine et l'image corrompue. 2ème ligne : le résultat de restauration par régularisation et par fonction harmonique.

2.4 Méthodes spectrales pour la classification

☞ Publications associées : [RI7, RI8,RI9, CI6, CI12, CN8, CN9, CN10, CN11, CO2, CO4] [ThèseE1] [StageR7] [StageR6][StageA2]

Dans les sections précédentes, nous avons présenté une interprétation des Laplaciens via des opérateurs en utilisant des versions discrètes de ce problème. Dans cette section, nous présentons une autre approche matricielle basée sur l'analyse spectrale de versions normalisées du Laplacien où le filtrage et les modèles de diffusions sont basés sur des matrices de Markov. Tout d'abord, nous rappelons le lien entre la diffusion spectrale basée sur les marches aléatoires sur graphe et la diffusion p-Laplacienne (approche fonctionnelle). Nous introduisons ensuite notre approche de caractérisation spectrale basée sur les marches aléatoires et nous montrons comment exploiter la non localité et le spectre issu de cette matrice pour caractériser la saillance visuelle et classer des objets.

2.4.1 Introduction et problématique

De nombreuses méthodes basées sur des graphes ont été proposées dans le domaine de l'analyse, du traitement et de la classification de données. Elles sont basées sur des concepts provenant de la théorie des graphes et de l'analyse spectrale et exploitent les propriétés spectrales du Laplacien ainsi que les modèles de diffusion associés. Ces méthodes sont devenues très populaires pour de nombreuses applications telles que la réduction de dimension, le regroupement de données similaires ou la classification.

Dans un grand nombre d'applications, nous sommes souvent confrontés à des données de grandes dimensions. Cette grande dimension représente une entrave pour le traitement, l'organisation, la recherche, l'analyse ou la visualisation de ces données. Ces problèmes sont classiquement abordés par les techniques de sélection de variables et de réduction de dimension, qui visent à trouver des structures intrinsèques de dimension réduite. Traditionnellement, cette réduction de la dimension est réalisée par des techniques linéaires telles que l'Analyse en Composantes Principales (ACP) [Pea01] ou l'analyse factorielle [Spe04]. Cependant, ces techniques ne peuvent pas traiter correctement des données complexes comme les techniques non linéaires [Bur10][SWH+06][TSL00].

A partir d'un ensemble de données $X = \{x_1, x_2, \cdots, x_N\}$ de N vecteurs de dimension D avec $x_i \in \mathcal{R}^D$, on désire trouver un nouvel ensemble $Y = \{y_1, y_2, \cdots, y_N\}$ de N vecteurs de dimension d avec $y_i \in \mathcal{R}^d$ avec $d \ll D$. On cherche donc à avoir $\left\|y_i - y_j\right\|_2$ qui soit faible lorsque x_i et x_j sont proches.

La figure 2.20 montre une taxonomie des techniques de réduction de la dimension [vdMPvdH07]. Toutes ces méthodes de réduction de données exploitent le contenu spectral d'une matrice de similarité mesurant la distance entre les données deux à deux. Les vecteurs propres fournissent une représentation dans un espace où les données sont mieux différenciées.

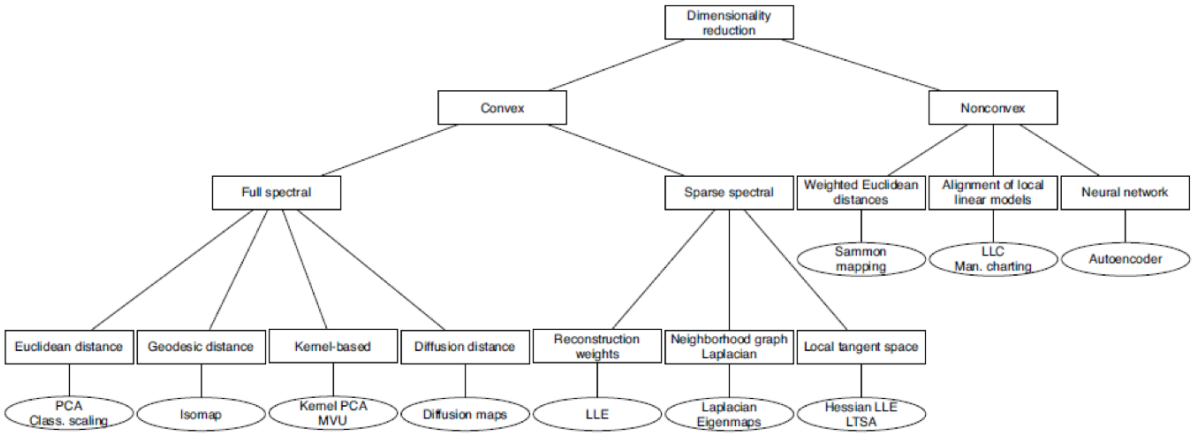


FIGURE 2.20: Aperçu des méthodes de réduction de données.

La figure 2.21 montre la réduction de dimension obtenue pour un ensemble de points représentant une structure non linéaire discrète (*SwissRoll*). On remarque que la projection effectuée respecte bien la géométrie initiale des points.

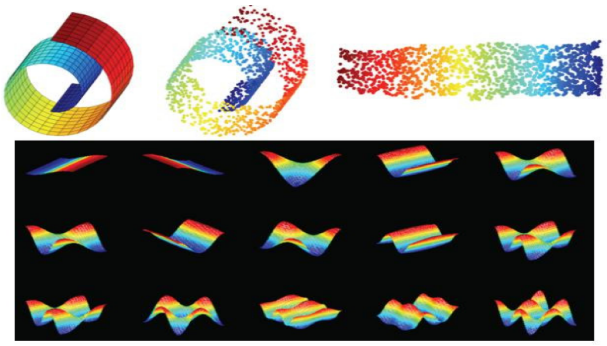


FIGURE 2.21: Réduction de dimension sur des points provenant d'une distribution de type "swiss-roll" (1ère ligne) et les vecteurs propres obtenus (Les lignes suivantes).

La méthode de regroupement spectral consiste à extraire les vecteurs propres associés aux plus grandes valeurs propres d'une matrice d'affinité normalisée. Ces vecteurs propres constituent un espace de dimension réduite dans lequel les données transformées seront linéairement séparables. Deux principales classes d'algorithmes ont été développées à partir de partitionnement de graphes.
La première classe est fondée sur un partitionnement bipartite récursif à partir du vecteur propre associé à la seconde plus grande valeur propre du graphe du Laplacien normalisé [KVV04] [SM00], ou vecteur de Fiedler [Chu97] dans le cas non-normalisé.
La deuxième classe d'algorithmes n'utilise pas de manière récursive d'un seul vecteur propre mais propose de projeter les données originales dans un espace défini par les k plus grands vecteurs propres d'une matrice d'adjacence normalisée (ou matrice similaire à celle-ci), et d'appliquer un algorithme standard comme k-means sur ces nouvelles coordonnées [MS01].

C'est dans l'esprit de cette dernière classe, que nous avons développé nos méthodes de classification basées sur les marches aléatoires sur graphe dans un souci de coût numérique et de simplicité algorithmique.

2.4.2 Marches aléatoires sur graphe

Les marches aléatoires ont été popularisées par le PageRank de Google. Nous proposons ici une autre manière de définir un ordre de données (issues des séquences d'images). Nous cherchons cette fois à définir l'ordre par une réduction non linéaire de dimension grâce à des marches aléatoires sur graphes.

2.4.2.1 Principe

Soit $G = (V, E, w)$ un graphe construit à partir de l'ensemble des données X et nous allons nous intéresser à un processus de marche aléatoire (ou de diffusion dans le graphe G).

Une marche aléatoire sur un graphe est un processus stochastique qui parcourt le graphe en sautant aléatoirement de sommet en sommet. La probabilité de transition du sommet u vers le sommet v est définie à chaque étape par :

$$p_{uv} = p(u, v) = p(x_u, x_v) = \frac{w(u, v)}{deg(u)} \tag{2.38}$$

On notera P la matrice de transition associée qui est alors définie par $P = D^{-1}W$ où D est la matrice des degrés des noeuds et W la matrice des poids associée au graphe. Cette matrice est aussi appelée matrice de diffusion ou matrice de propagation. Ainsi, pour toute fonction f définie sur le graphe, on peut écrire : $Pf(u) = \sum_{v \in \mathcal{N}(u)} p(u, v) f(v)$.

$P = D^{-1}W$ n'est pas symétrique mais peut être ré-écrite sous une autre forme : $S = D^{-\frac{1}{2}} P D^{\frac{1}{2}} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. P et S ont les mêmes valeurs propres.
$L_n = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = I - S$
Cette écriture du Laplacien normalisé a l'avantage de montrer clairement le lien entre le Laplacien et les marches aléatoires. Il existe une connexion forte entre les marches aléatoires et le regroupement spectral. En effet, si λ_i est une valeur propre de L_n alors $1 - \lambda_{n-(i-1)}$ est une valeur propre de P.

Considérons le noyau $p^t(u, v)$ correspondant à la t^{ieme} puissance de P, qui peut être interprété comme la probabilité pour un marcheur d'atteindre le sommet v en partant du sommet u en t étapes. La matrice de transition de la chaîne de Markov correspondante est $P^t = (p^t(u, v))$. Les états ici sont les sommets du graphe. La matrice P^t est stochastique ($\forall u, \forall v \ 0 \leqslant p^t_{uv} \leqslant 1$ et $\sum_{u \in V} p^t(u, v) = 1$).

Si l'on veut décrire la probabilité de transition $p^t(u, v)$ d'un noeud u à un noeud v en t étapes, il suffit de considérer des voisinages plus larges, ce qui correspond à élever la matrice P à la puissance t. Comme le graphe est connexe et non bipartite, alors la marche aléatoire converge vers une distribution stationnaire unique $\pi = [\pi_1, \cdots, \pi_n]$

satisfaisant $P^t\pi = \pi$ (i.e. $\lim_{t\to\infty} p^t(u,v) = \pi_0 = \phi_0(v)$ avec $\phi_0(u) = \frac{deg(u)}{Vol(V)}$

π est un vecteur qui traduit une distribution statistique de l'occupation des états de la chaîne de Markov correspondant à la marche aléatoire. Les valeurs propres de P^t sont : $\lambda_1^t = 1 \geqslant \lambda_2^t \geqslant \cdots \lambda_N^t \geqslant -1$. Soit $\{\psi_l(v)\}$ les vecteurs propres de P^t. $\forall v \in V$ et $1 \leqslant l \leqslant N$.

2.4.2.2 Distance de diffusion

Il a été démontré [LL06] que les distances de diffusion dans l'espace original sont égales aux distances euclidiennes dans l'espace de projection de diffusion. Ce résultat justifie l'utilisation des distances euclidiennes dans l'espace de projection de diffusion pour le clustering.

La distance de diffusion $D_t(u,v)$ entre deux noeuds u et v peut être calculée à partir des probabilités de transitions par :

$$D_t^2(u,v) = \sum_{l\in V}(p^t(u,l) - (p^t(v,l))^2 \tag{2.39}$$

Elle peut être également exprimée en termes de valeurs et vecteurs propres de P^t par :

$$D_t^2(u,v) = \sum_{l\geqslant 1}(\lambda_l^t\psi_l(u) - \lambda_l^t\psi_l(v))^2 \tag{2.40}$$

À partir de l'équation (2.40), nous pouvons remarquer qu'on peut réduire la dimension en négligeant certaines dimensions dans l'espace de diffusion. L'ensemble des données $\{x_i\}_{i=1\cdots N}$ peut être représenté par :

$$y_i = \psi_t(x_i) = \begin{pmatrix} \lambda_0\psi_0^t(x_i) \\ \lambda_1\psi_1(x_i) \\ \vdots \\ \lambda_{d-1}\psi_{d-1}(x_i) \end{pmatrix} \tag{2.41}$$

avec $d \ll N$ et $\psi_1(x_i)$ indique le i_{eme} élément du 1er vecteur propre de P. Le but est de représenter chaque $x \in \mathbb{R}^N$ par un point $y_i \in \mathbb{R}^d$, avec $d \ll N$, de sorte que l'ensemble $Y = \{y_1, y_2, \cdots, y_d\}$ capture toute l'information géométrique intrinsèque de l'ensemble d'origine.

Notons que la chaîne de Markov (à densité invariable) peut être construite après normalisation du noyau comme suit : $\widetilde{w}\ (u,v) = \frac{w(u,v)}{deg(u)deg(v)}$. Ainsi la nouvelle probabilité de transition devient :
$p(u,v) = \frac{\widetilde{w}(u,v))}{\widetilde{deg}(u)}$ avec $\widetilde{deg}\ (u) = \sum_{v\sim u} \widetilde{w}\ (u,v)$

La première normalisation de la matrice de similarité permet de trouver une représentation indépendante de la distribution. La seconde sert à rendre le noyau stochastique.

Notons que le premier vecteur propre est constant et que l'ordre dans ψ_2 fournit une information sur la structure en groupes du graphe. Ce second vecteur propre ψ_2 est connu comme le vecteur de Fiedler et peut être utilisé pour ordonner l'ensemble des données X. Belkin et Niyogi [BN03] ont montré que l'utilisation de l'opérateur de Laplace-Beltrami constitue un choix intéressant pour la prise en compte de l'information géométrique. En particulier, l'information locale est bien préservée par l'utilisation de ces fonctions propres. L'approximation de l'opérateur Laplace- Beltrami est présenté dans l'algorithme complet ci-dessous :

Algorithme 1 Réduction de dimension par marches aléatoires sur graphe

Entrées: $X = \{x_1, x_2, \cdots, x_n\} \subset \mathbb{R}^N, d, \epsilon, \alpha$
Sorties: $Y = \{y_1, y_2, \cdots, y_d\} \subset \mathbb{R}^d$

1. Construction de la matrice de similarité W en utilisant le noyau gaussien w_ϵ
2. Normalisation du noyau :
   $\widetilde{w}(x_i, x_j) = \frac{w(x_i, v_j)}{(d(v_i)d(v_j))^\alpha}$
   Remarque : Nous retrouvons des cas classiques pour certaines valeurs de α.
   - $\alpha = 0$: Graphe de Laplacien.
   - $\alpha = 0,5$: le propagateur Fokker-Plank.
   - $\alpha = 1$: l'opérateur de Laplace-Beltrami.
3. Construction de la matrice de transition P avec le noyau :
   $p(x_i, x_j) = \frac{\widetilde{w}(x_i, x_j))}{\widetilde{d}(x_i)}$ avec $\widetilde{d}(x_i) = \sum_{u \sim v_i} \widetilde{w}(u, v_i)$
4. Projection dans l'espace réduit et exploitation des distances de diffusion.
   Espace de diffusion : $\{\lambda_i \psi_i\}$
   $x \to y = (\lambda_1\psi_1(x_1), \lambda_2\psi_2(x_2), \cdots, \lambda_d\psi_d(x_d))$
   Distance de diffusion :
   $D_t^2(x_i, x_j) = \sum_{l \geqslant 0} \lambda_l^2(\psi_l(x_i) - \psi_l(x_j))^2 = \sum_{x_l \in V} (p(x_i, x_l) - (p(x_j, x_l))^2$ avec
   $1 = |\lambda_1^t| \geqslant |\lambda_2^t| \cdots \geqslant 0$

La figure 2.22 illustre la visualisation d'un ensemble de patchs centrés sur les points SIFT d'une image.

2.4.2.3 Diffusion et lien avec la régularisation

L'apprentissage semi-supervisé consiste, à partir d'un certain nombre d'échantillons de données possédant des étiquettes connues, à déterminer l'étiquette de nouveaux échantillons. Plusieurs techniques ont été proposées dans la littérature pour la diffusion de labels en utilisant les graphes, certaines sont basées sur l'approche spectrale [WWL07] [Azr07] et d'autres sont basées sur la minimisation de fonctionnelles [ZS04]et [WhLZH06].

Soit $\mathcal{L}$ l'ensemble des labels. Pour une classification en deux classes, on considère souvent $\mathcal{L} = \{+1, -1\}$ avec $y_u \in \mathcal{L}$ le label du sommet u. Soit y^t le vecteur de labels des sommets à l'itération t : $y^t = (y_1^t, \cdots, y_l^t, \cdots, y_N^t)^T$

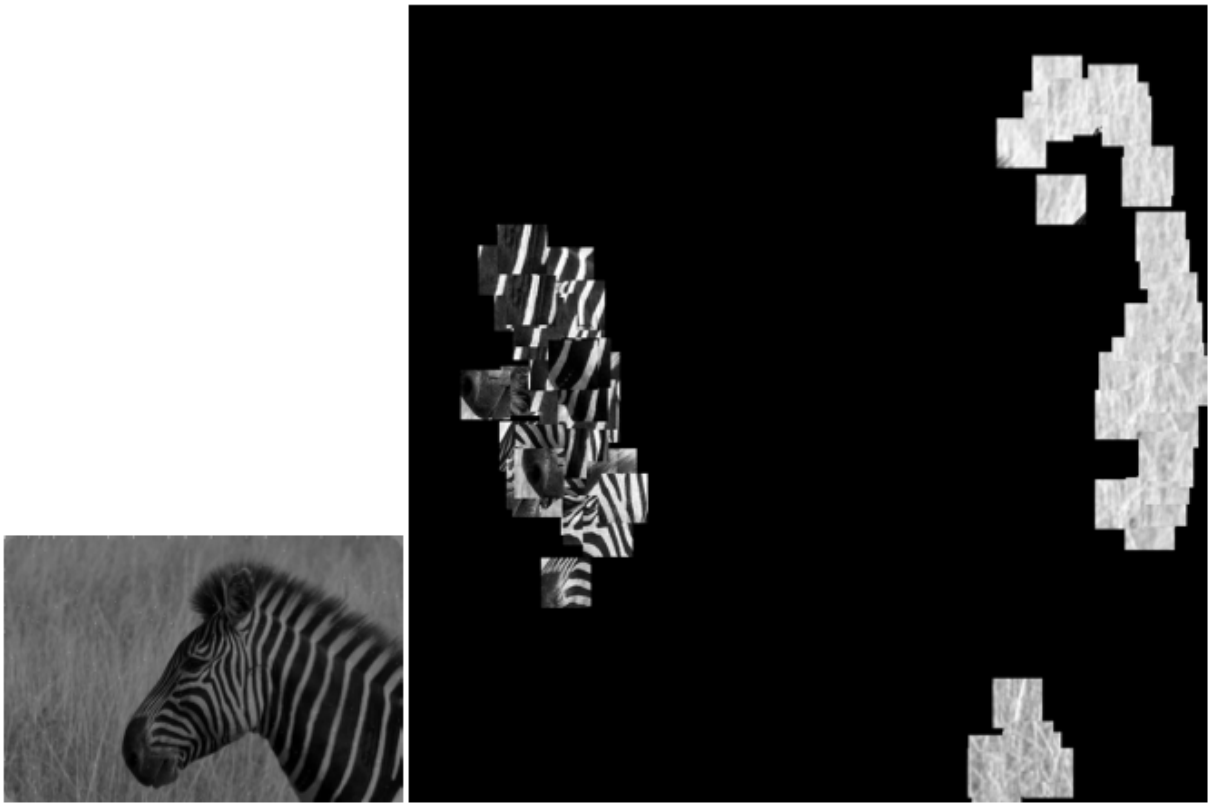


FIGURE 2.22: Projection des patchs 50 × 50 de l'image Zebra centrés sur les points SIFT dans l'espace réduit.

Nous avons proposé dans [**?**] une stratégie de propagation qui se fait de manière itérative :

$$y_v^{t+1} = \sum_{u \in \mathcal{N}(v)} p_{uv}^t y_u^t \tag{2.42}$$

avec v : le sommet courant.

y_v^{t+1} : le nouveau label du sommet à l'itération $t+1$.

p_{uv} : est la probabilité de transition du sommet u vers le sommet v.

y_u^t : le label du voisin u du sommet v à l'itération t.

En découpant la matrice de propagation P et y^t en parties étiquetées et non étiquetées : $y^t = ((y_L^t)^T, (y_U^t)^T)^T$, $P = \begin{pmatrix} P_{LL} & P_{LU} \\ P_{UL} & P_{UU} \end{pmatrix}$

La stratégie de propagation devient alors :

$$y_U^{t+1} = P_{UL} y_L + P_{UU} y_U^t \tag{2.43}$$

En réécrivant l'eq 2.21 sous forme matricielle, nous obtenons une régularisation spectrale (filtrage dans le domaine spectral) c.à.d. un filtrage par diffusion sur des graphes

basé sur des matrices de Markov. En effet :

$$f^{(t+1)}(v) = \frac{\lambda}{\lambda + d^{(t)}_{\gamma_f}(v)} f^0(v) + \sum_{u \sim v} \frac{\gamma^{(t)}_f(u,v)}{\lambda + d^{(t)}_{\gamma_f}(v)} f^t(u) \quad \forall v \in V \tag{2.44}$$

avec $d^{(t)}_{\gamma_f}(v) = \sum_{u \sim v} \gamma^{(t)}_f(u,v)$ et $\gamma^{(t)}_f(u,v) = w^{(t)}(u,v)(\left\|\nabla f^{(t)}(v)\right\|^{p-2} + \left\|\nabla f^{(t)}(u)\right\|^{p-2})$

On peut noter que lorsqu'il n'y a pas d'attache aux données ($\lambda = 0$) et quand le paramètre $p = 2$, la fonction f est la solution de la minimisation du problème de Dirichlet :

$$\sum_{u \in V} \left\|\nabla_w f(v)\right\|_2^2 \tag{2.45}$$

et devient :

$$f^{(t+1)}(v) = \sum_{u \sim v} \frac{w^{(t)}(u,v)}{\sum\limits_{u \sim v} w^{(t)}(u,v)} f^t(u) = \sum_{u \sim v} p^{(t)}(u,v) f^t(u) \quad \forall v \in V \tag{2.46}$$

En posant $y^t_v = f^{(t)}(v)$, on retrouve ainsi le processus de diffusion spectral (2.42). De même lorsque l'algorithme converge, la fonction f obtenue correspond à la solution de l'équation de la chaleur :

$$\Delta^i_{2,w} f(u) = \Delta_w f(u) = 0, \forall u \in V$$

La figure 2.23 illustre le résultat de propagation de labels sur deux images, l'une obtenue avec deux labels et l'autre avec trois labels.

FIGURE 2.23: De gauche à droite et de haut en bas : l'image originale, l'image marquée avec deux labels (resp. trois labels) et la carte des labels.

Nous avons également, dans le cadre d'un travail préliminaire [**StageR2**] étendu ce travail aux vidéos. On peut observer la propagation des labels alors que les séquences sont en mouvement. La figure 2.24 montre le résultat de propagation de labels sur la séquence vidéo d'Akiyo.

FIGURE 2.24: Diffusion interactive de labels sur une séquence vidéo. De gauche à droite : la frame 42 labélisée par l'utilisateur, le même frame au bout de 100 itérations de la propagation de labels et enfin la frame 32 au bout de 100 itérations.

2.4.3 Applications

Nous avons proposé des solutions basées sur les marches aléatoires sur graphes pour différentes applications :

Caractérisation de la saillance visuelle par l'énergie de la matrice de transition (Markov) P correspondant au graphe de diffusion locale.

Nous associons à chaque pixel centre d'un patch de taille donnée une fenêtre de forme quelconque (carrée, ligne, colonne, $\cdots$). Le graphe est formé à partir de l'ensemble des noeuds de la fenêtre. La similarité entre les noeuds est calculée par la formule 2.24. Le descripteur que nous avons proposé est calculé à partir de la trace de la matrice de diffusion associée au graphe. Il constitue une mesure qui reflète la connectivité de l'ensemble des noeuds du graphe local.

Soit Γ_u le descripteur caractérisant un noeud u formé de la somme des valeurs propres.

$$\Gamma_u = \sum_{h=1}^{\mathcal{N}(u)} \lambda_h = Tr(P) - 1 \tag{2.47}$$

Le descripteur peut être calculé sans passer par la décomposition spectrale de la matrice de transition P. Ceci offre le grand avantage de se défaire du temps de calcul que peut présenter la décomposition spectrale.

Pour illustrer les résultats de la diffusion géométrique locale, nous avons appliqué l'algorithme sur des images synthétiques (voir figure 2.25). Les valeurs singulières fournissent l'information sur la dépendance entre les lignes et les colonnes de la matrice de transition locale correspondant à chaque pixel. Comme on peut le constater, le spectre de la matrice de Markov locale reflète bien les variations lumineuses entre les pixels le long de la fenêtre.

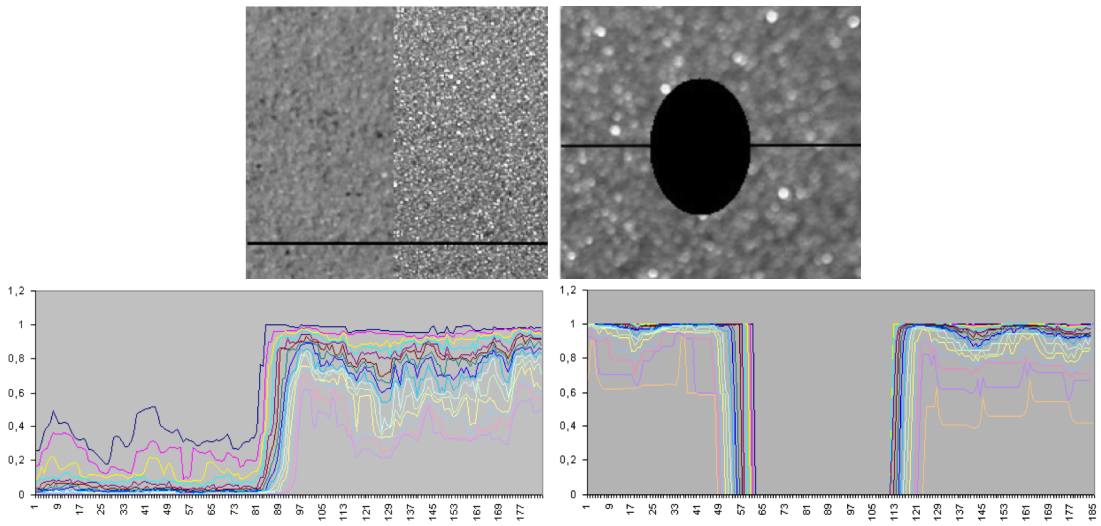


FIGURE 2.25: En haut : Textures de taille 200×200. En bas : Spectre ($\lambda_1 \cdots \lambda_{16}$) de la matrice du graphe de transition le long de la ligne. Paramètres : Un pixel dans un voisinage de 16×16

.

La figure 2.26 montre des résultats de l'utilisation de ce descripteur dans un processus de caractérisation et de segmentation d'objets visuels. La figure présente également une comparaison avec le descripteur proposé par [THES06] pour caractériser la texture et qui est composé de la somme des valeurs propres de la matrice carrée centrée autour de chaque pixel. Nous avons utilisé ce descripteur dans divers applications de segmentation d'objets visuels. Les pixels qui possèdent une fréquence élevée du spectre correspondent aux objets d'intérêt. La figure 2.27 présente un exemple de caractérisation et de localisation des composantes faciales.

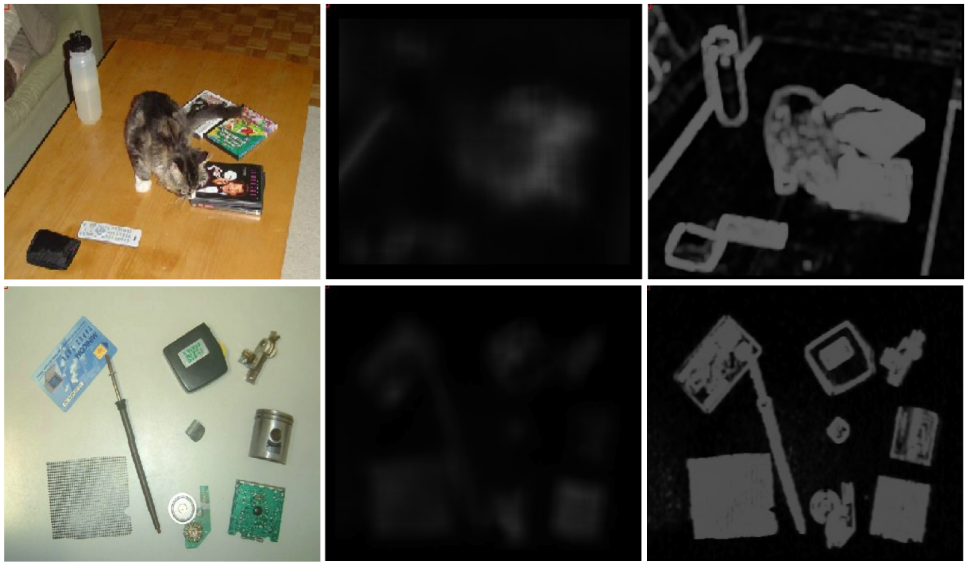


FIGURE 2.26: Saillance visuelle : De gauche à droite : l'image initiale, l'approche spectrale de Targhi et notre approche avec un patch 5×5 et une fenêtre 25 × 25

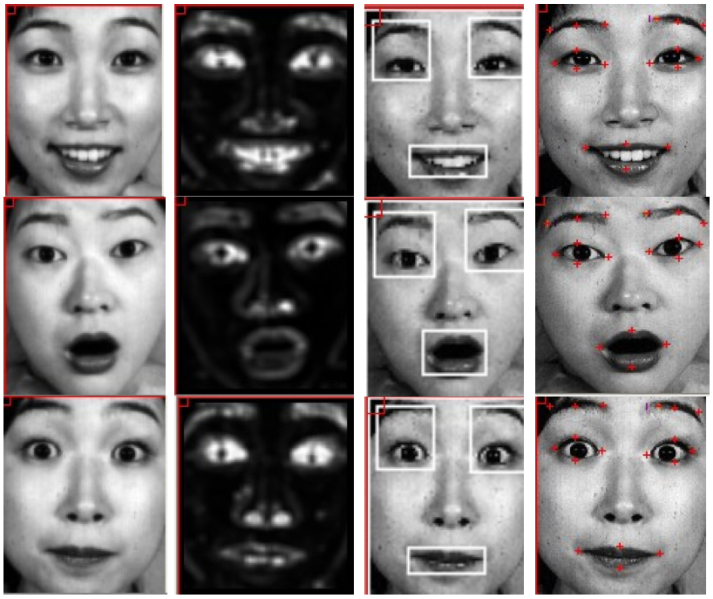


FIGURE 2.27: Utilisation de la saillance dans un processus d'extraction de composantes faciales.

La figure 2.28 montre un autre exemple d'utilisation du descripteur pour segmenter des images d'algues, et une comparaison avec la méthode basée sur les coupes normalisées proposée par [SM00].

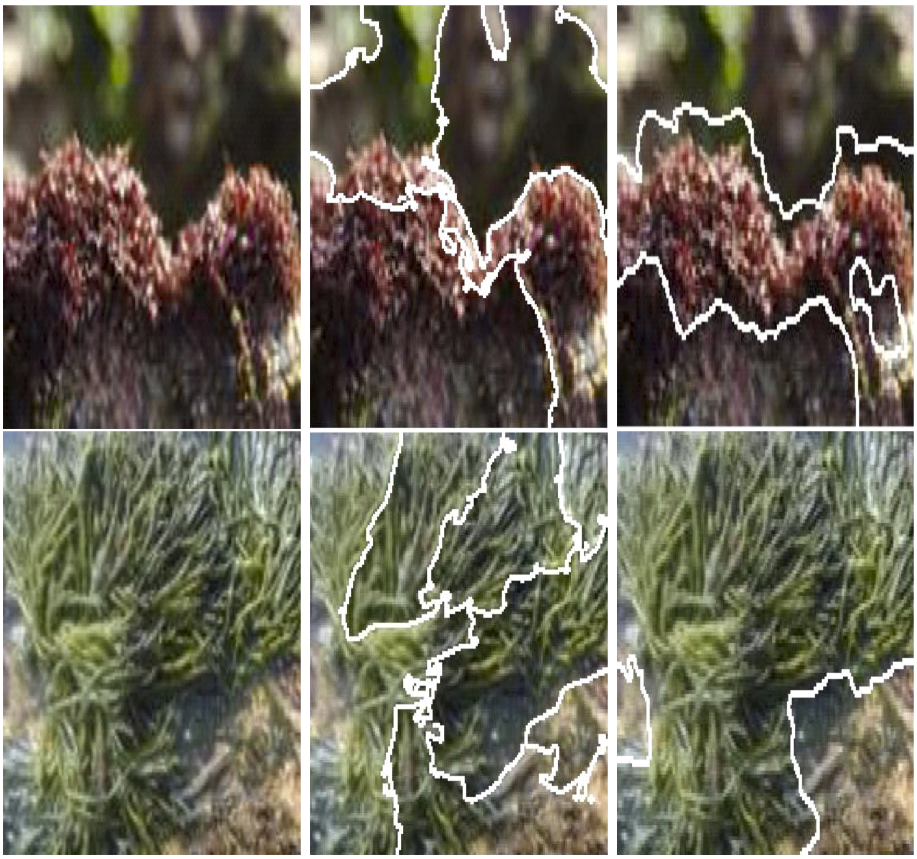


FIGURE 2.28: **A gauche** : Images d'algues. **Au milieu**, résultat de segmentation par coupes normalisées et **à droite** le résultat de notre approche.

Segmentation d'image par diffusion géométrique globale des points caractéristiques sur graphe : Le principe consiste à construire un graphe à partir d'un nombre réduit de sommets (germes) répartis entre le fond et l'objet d'intérêt et de les classer après réduction de la dimension. Une fois les germes classés, nous utilisons la méthode de croissance de régions (lignes de partage des eaux) pour segmenter les objets d'intérêt.

La figure 2.29 illustre les différentes étapes de l'algorithme. D'autres résultats de segmentation sont présentés à la figure 2.30.

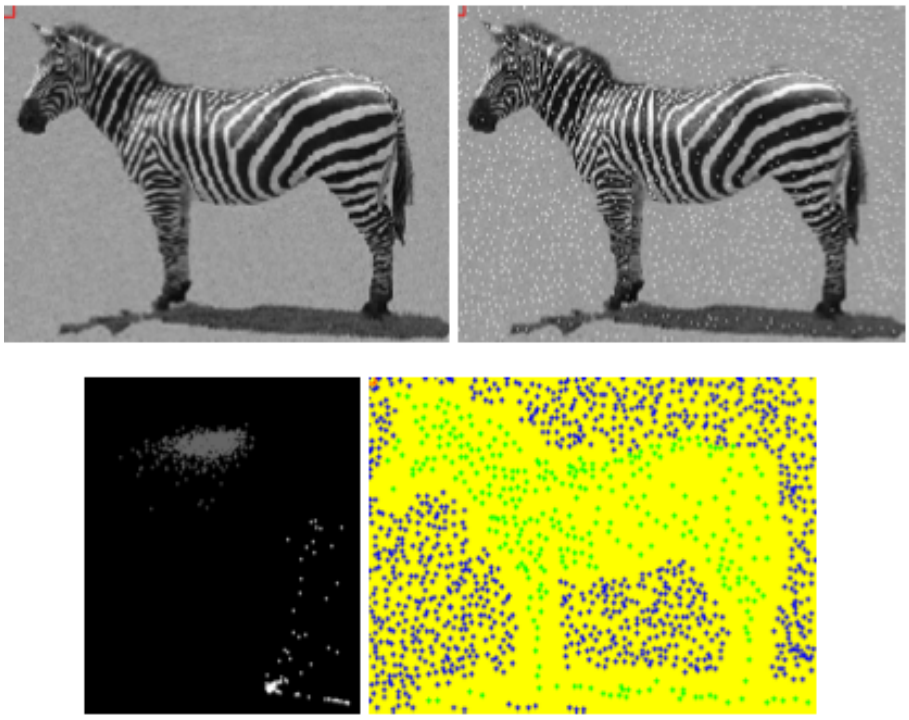


FIGURE 2.29: 1ère ligne : Image et points caractéristiques. 2ème ligne : Projection des points sur les deux axes principaux $\{\lambda_2\psi_2\}$ et $\{\lambda_3\psi_3\}$ classification par k-means.

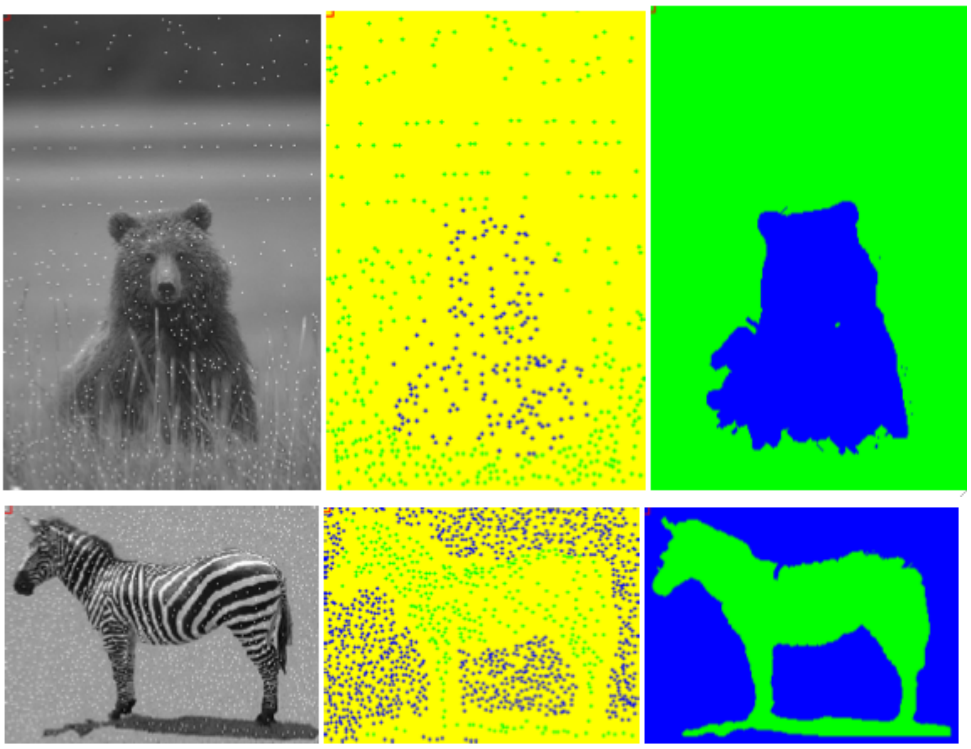


FIGURE 2.30: Classification dans l'espace spectral. De gauche à droite l'image d'origine, les extrema, la projection dans l'espace réduit et classification par k-means.

Bouclage de pertinence par diffusion sur graphe. Le problème de bouclage de pertinence est formulé comme un problème d'apprentissage semi-supervisé. Nous avons proposé une stratégie d'affinage basée sur la propagation de labels (cf 2.42) dans le cadre de la recherche interactive dans une base de 2000 images d'algues. L'idée vise à construire une fonction de pertinence f_y à valeurs dans [-1, 1] sur un ensemble X en fonction d'un ensemble d'apprentissage y. Dans notre cas, l'ensemble X à classifier sont les n images de la base d'algues. Cela permet de distinguer trois cas différents : l'appartenance à la catégorie recherchée (proche de 1), la non appartenance à la catégorie recherchée (proche de -1), et l'incertitude (proche de 0). Pour mettre en évidence l'apport du bouclage par diffusion, nous avons sélectionné 40 images qui couvrent l'ensemble des algues pour être utilisées comme requêtes. Pour chaque requête, nous avons évalué les 16 premières images retrouvées. L'idéal serait que les 16 premières appartiennent à la même espèce que l'image de requête. La performance est mesurée par [MM96] :

$$TauxR = \frac{\text{Nombre d'images pertinentes}}{\text{Taille de la classe}} \times 100 \tag{2.48}$$

Nous avons obtenu une amélioration du taux de récupération moyenne de 70.26% à la première itération jusqu'à 93,54% au bout de la 3ème itération. Nous avons évalué la performance globale sur 4 espèces d'algues en termes de taux de précision et de rappel après trois itérations. Les résultats obtenus ont montré l'apport significatif du bouclage de pertinence par diffusion dans l'amélioration de la précision et du rappel (voir figure 2.31).

2.4.4 Conclusion et discussion

Dans cette section, nous avons proposé une autre manière de définir un ordre de données par une réduction non linéaire de dimension grâce à des marches aléatoires sur graphes. L'utilisation des premiers vecteurs propres du Laplacien de graphe comme espace de projection dans lequel les clusters sont constitués est en général justifiée comme une relaxe du problème discret de clustering. Nous avons exploité le lien entre le Laplacien du graphe et les chaînes de Markov pour proposer des méthodes de classification, avec des matrices de similarité adaptées, et des techniques de classification appropriées. Nous avons présenté nos trois contributions qui sont basées sur les marches aléatoires sur graphe :

1. Une méthode de classification par réduction de la dimension de la matrice de diffusion globale. Il s'agit d'une interprétation probabiliste du regroupement spectral basée sur un modèle de diffusion global où la distance de diffusion utilisée est une distance basée sur une marche aléatoire sur le graphe. La projection de diffusion de l'espace des données dans un espace est définie par les k premiers vecteurs propres.
2. Un descripteur de saillance visuel basé sur la trace de la matrice de transition locale.
3. Un processus de diffusion de labels pour l'apprentissage semi-supervisé en utilisant la matrice de propagation.

Pour la première contribution, nous avons présenté une méthode de segmentation d'image à partir d'un ensemble de points caractéristiques. L'idée est de projeter cet ensemble dans

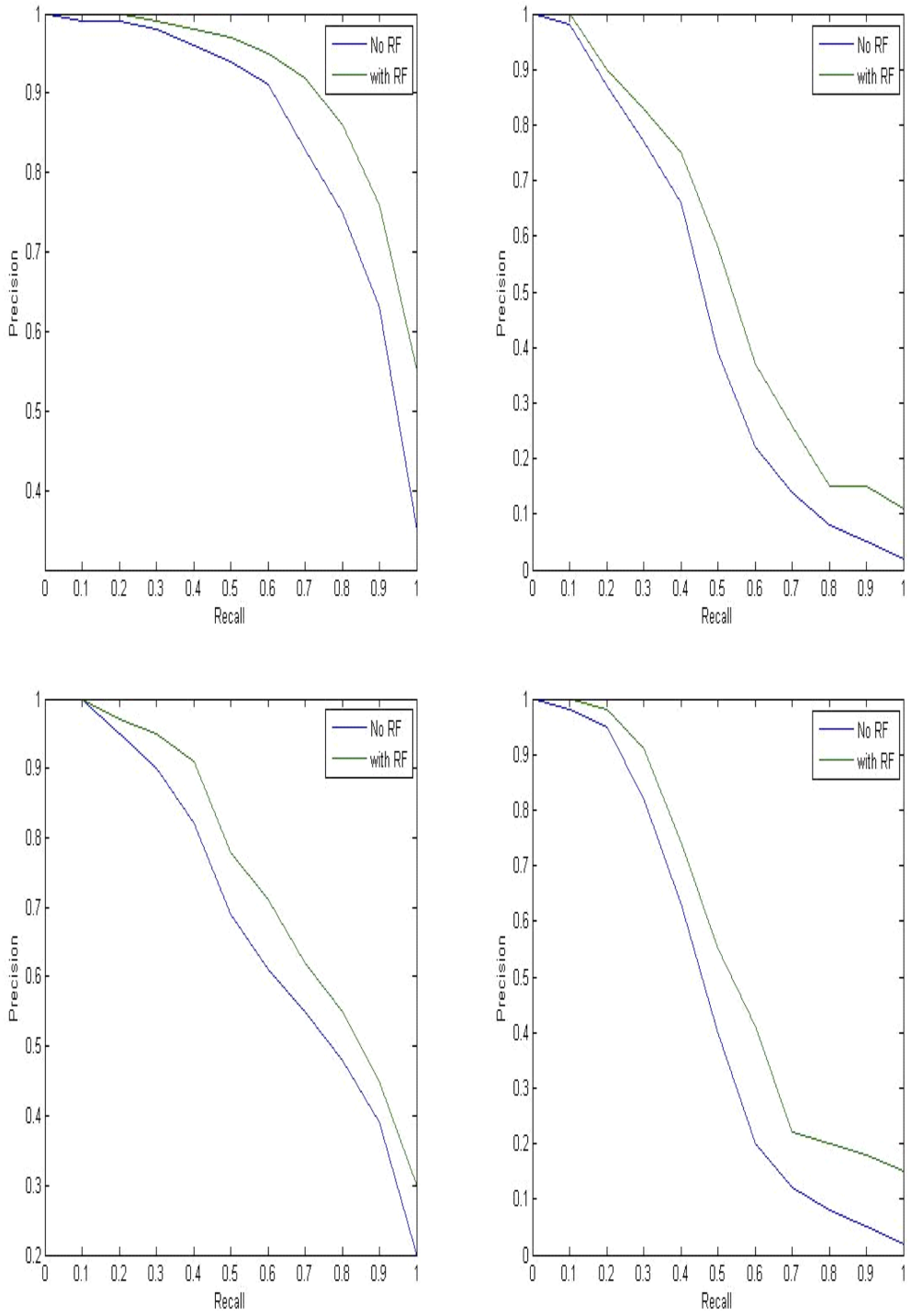


FIGURE 2.31: Taux de précision et de rappel avec et sans la diffusion sur une base de 4 espèces d'algues différentes (a) Gelidium sesquipedale, (b) Codium fragile, (c) Fucus spiralis et (d) Blidinga minima.

l'espace de diffusion réduit, de faire la classification et de revenir à l'espace d'origine pour

segmenter les objets d'intérêt en utilisant une méthode morphologique basée sur la LPE.

Pour la seconde contribution, nous avons présenté une application de ce descripteur pour la caractérisation d'objets visuels et de la texture.

Enfin, nous avons proposé une application du processus de diffusion pour formuler le problème de bouclage de pertinence sur une base d'images maritimes.

Cette approche matricielle repose sur la seule mesure de transition entre tous les noeuds du graphe construit, sans a priori sur les formes des classes (ou clusters). Le paramètre de l'affinité gaussienne σ joue un rôle crucial dans le partitionnement des données. L'estimation d'un ordre de grandeur est possible mais la recherche d'un optimum reste un problème ouvert.

3

Analyse de séquences vidéo

Sommaire

Résumé

Je présente dans ce chapitre mes travaux en analyse d'enregistrements vidéo de gestes humains. La première partie traite de l'adaptation de méthodes basées sur les arbres de décision pour la recherche d'attributs pertinents permettant de caractériser les pixels de peau. Le second axe concerne le développement d'une méthode de segmentation convexe binaire d'un individu en mouvement intégrant les contours actifs et le flot optique. Un troisième axe traite de la catégorisation de bases de gestes et d'actions en utilisant la régularisation discrète sur graphe et les moments spatio-temporels comme descripteurs des vidéos. La dernière partie traite d'une solution pratique d'un système de suivi et de la prédiction du regard par processus gaussiens.

Pour observer et documenter un événement en vidéo tel que le comportement d'un objet de référence, il est important de *segmenter* et *suivre* cet objet avant de *classer* et de *reconnaître* l'événement qui s'y rapporte. Les applications se distinguent par le type d'objets que l'on souhaite suivre, mais également par l'objectif de l'analyse. Cette variabilité des besoins explique pourquoi, jusqu'à présent, la plupart des solutions informatiques proposées pour l'analyse de séquences vidéo ont un caractère dédié. Dans ce chapitre, notre attention est focalisée sur l'étude de gestes et actions prédéfinis d'un individu en mouvement.

Ce chapitre présente les recherches que j'ai menées en analyse vidéo et en apprentissage automatique. Une première partie, la plus ancienne de mes travaux, a été consacrée à l'extraction des attributs pertinents concernant la couleur de la peau en utilisant les outils de datamining. Ces travaux ont été réalisés essentiellement avec Liming Chen dans le cadre de la thèse de M. Hammami que nous avons co-encadrée.

Ensuite, je présente la deuxième partie de mes travaux liés à la segmentation d'un objet en mouvement par contours actifs. Ce travail a été réalisé avec Abderrahim Elmoataz dans le cadre de la thèse co-encadrée de Y. Zinbi. La dernière partie des travaux portant sur

l'analyse des gestes et du regard ont été essentiellement réalisés avec F. Jouen dans le cadre des masters de B. Safadi et de la thèse de N-B. Linh que nous avons co-encadrés. D'autres travaux liés à la problématique d'analyse vidéo ont été effectués dans le cadre de stages de masters de F. Haddad, Y. Lan, M. Maameri et de ma collaboration avec I. Lassoued.

Mes travaux présentés dans ce chapitre s'inscrivent dans une démarche d'analyse de séquences vidéos de gestes et d'actions qui consiste à :

Apprendre pour caractériser : Utiliser les outils de fouille de données pour une analyse colorimétrique et spectrale afin de trouver les attributs pertinents caractéristiques de la couleur de peau.

Segmenter et suivre un individu en mouvement par une méthode convexe adaptée basée sur les contours actifs et le flot optique.

Classer et reconnaître des gestes et actions d'un individu par SVM en se basant sur les moments de Zernike. La classification peut être améliorée par la régularisation discrète sur graphe.

Prédire à partir d'un ensemble d'observations (vérité terrain) et selon une stratégie adaptée (comme le paradigme gaussien). A partir d'une vérité terrain provenant de tests occulométriques, nous avons proposé une méthode de prédiction du regard par processus gaussiens.

Dans chacune des sections, je commencerai par décrire le contexte et l'état de l'art puis je donnerai une description synthétique de mes contributions. Le plan du chapitre est organisé comme suit :

- Dans la section 3.1, nous présentons nos attributs jugés pertinents pour caractériser la couleur de la peau dans une image et les résultats obtenus.
- Dans la section 3.2, nous proposons une approche de segmentation convexe d'un objet en mouvement basée sur un modèle basé région des contours actifs. Nous présentons notre modèle basé sur la densité de probabilité pour la segmentation en volume d'un objet d'intérêt (multi-régions). Le mouvement de l'objet d'intérêt est pris en compte via un critère supplémentaire issu du calcul du flot optique.
- Dans la section 3.3, nous exploitons des structures basées sur les graphes pondérés pour représenter l'ensemble des vidéos de gestes et d'actions. où les sommets sont représentés par les volumes binaires avec des descripteurs basés sur les moments spatio-temporels de Zernike. Nous présentons nos résultats de classification par SVM et nous montrons l'apport de la régularisation discrète sur graphe (réduction de dimension et débruitage), construit dans l'espace de faible dimension, pour l'amélioration des résultats de la classification.
- Dans la section 3.4, nous présentons un système de capture et de prédiction du regard par processus gaussiens.

3.1 Espace hybride de couleur peau

☞ Publications associées : [RI12, RI13, RN2, CL1, CI16, CI18, CI19, CI20] [ThéseF1]

3.1.1 Contexte et problématique

Dans le cadre de la thèse de M. Hammami, nous nous sommes intéressés à la caractérisation de la couleur de la peau. Ce travail a émergé à travers le projet scientifique international COCO [1] pour le développement d'une plate forme adaptable pour l'indexation, l'édition, la production et la distribution intelligente de documents multimédias. C'est dans ce contexte que, nous avons développé un modèle de peau et mis au point un logiciel, appelé WebGuard, pour le filtrage sémantique du contenu Web.

Il s'agit de trouver les attributs pertinents pour caractériser les pixels de peau. L'objectif étant d'améliorer l'approche utilisée par Compaq [JR02a] qui est la méthode de référence. Nous cherchons donc à identifier un pixel de peau avec un degré élevé de précision en utilisant des règles de prédiction en fonction de différents espaces de couleur. Le modèle de peau recherché sera utilisé comme un traitement préliminaire dans plusieurs problèmes de reconnaissances des diverses parties du corps humain, que nous verrons dans ce chapitre, tel que les mains, le visage et toutes les parties nues du corps.

Les méthodes de détection de la couleur de la peau peuvent être classées en trois catégories :

Les approches non paramétriques : Ces méthodes déterminent les distributions des classes en ne se basant que sur les observations issues des différents échantillons des classes sans introduire aucune hypothèse sur les formes des distributions. La distribution de la couleur de peau est modélisée généralement par un histogramme. Ceci a l'avantage de ne pas faire de supposition sur le type de distribution. On peut ensuite calculer la probabilité qu'un pixel donné soit un pixel de peau. Le temps de calcul nécessaire est plus faible que pour les méthodes paramétriques. [JR02a][GSS02][zar99].

Les approches paramétriques : la distribution de la couleur de peau est modélisée par un mélange de gaussiennes. Une étape d'apprentissage permet de calculer les paramètres des gaussiennes, et ainsi de calculer la probabilité qu'un pixel donné soit un pixel de peau. Cette catégorie suppose que la distribution de couleur de peau puisse être modélisée par une gaussienne, ce qui n'est pas forcément évident. Le mélange de gaussiennes peut par exemple être estimé avec l'algorithme EM. Plusieurs travaux sur la modélisation de la distribution de la couleur de peau ont utilisé un mélange de gaussiennes [MSP03][HAMJ02][YKA02].

Les méthodes explicites utilisant des règles de décision empiriques et/ou statistiques pour la détection des pixels ayant la couleur de la peau [CW97] [CG99]. L'avantage de ces méthodes réside dans la simplicité des règles de détection de la peau qu'elles utilisent, ce qui permet une classification rapide. Cependant, leur problème principal est la difficulté de déterminer empiriquement un espace couleur approprié ainsi que des règles de décision adéquates qui assurent un taux de reconnaissance élevé.

1. Projet de coopération franco tunisien, "Courtiers Coopérants pour des services de qualité sur Internet".

3.1.2 Espace hybride de couleur peau

3.1.2.1 Approche bayésienne

Nous avons démarré notre étude par une approche bayésienne. Afin de réduire la complexité de l'étude, nous nous sommes borné à l'utilisation de deux axes pour la caractérisation de la couleur de peau, notre intuition étant que deux axes suffisent à discriminer les couleurs de peau de celles de non peau. Conformément à une approche bayésienne, nous avons donc construit des histogrammes de couleur de peau et de non-peau selon différentes combinaisons d'axes de couleur (issus des espaces de couleur RGB, rgb, HSV, YIQ, YCbCr,CMY) afin de déduire par la suite les combinaisons pertinentes qui représentent le mieux la distribution des couleurs de pixels de peau. La probabilité $Pr(peau/C_1C_2)$ est donnée par la formule de Bayes suivante :

$$Pr(peau/C_1C_2) = \frac{Pr(C_1C_2/peau.Pr(peau)}{Pr(C_1C_2/peau.Pr(peau) + Pr(C_1C_2/non\_peau.Pr(1 - Pr(peau))} \tag{3.1}$$

Mais cette approche a vite trouvé ses limites compte tenu de l'importance de la combinaison d'axes. En plus, l'apprentissage sur des données issues de probabilités sur une combinaison d'axes ne permet pas de discriminer le poids de chaque axe. Cette approche est intéressante mais elle ne nous a pas permis d'améliorer la performance de la méthode développée par Compaq (cf. 3.6). Cela nous a amené à exploiter directement les valeurs des pixels issues des différents axes de représentation afin de déterminer la pertinence de chaque axe indépendamment les uns des autres et d'en extraire les règles de prédiction.

3.1.2.2 Spectre de la lumière visible

Outre les espaces de couleur traditionnels utilisés auparavant nous avons pris en compte aussi le spectre de la lumière visible. Soit le spectre $Dist = M_f$ calculé pour chaque pixel :

$$M_f = r_0 \times M_R + v_0 \times M_V + b_0 \times M_B \tag{3.2}$$

où M_R, M_V, M_B sont respectivement les composantes rouge (R), vert (G) et bleu (B) d'un pixel, avec $r_0 = 700$, $v_0 = 546.1$, $b_0 = 435.8$ les composante primaires dans le système C.I.E.

Les expérimentations que nous avons conduites montrent qu'un pixel de peau est généralement caractérisé par une bande spectrale de longueur d'onde comprise entre 568 nm et 680 nm. Cet intervalle caractérise en réalité trois bandes spectrales qui sont le orange, le jaune et le rouge (cf. figure 3.1).

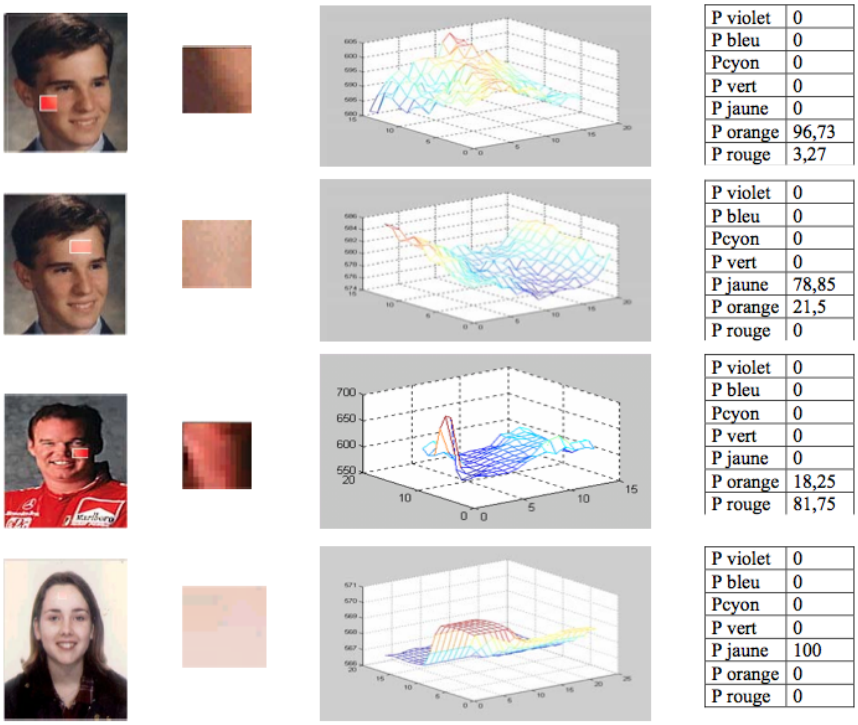


FIGURE 3.1: Spectres associés aux pixels de peau

3.1.2.3 Extraction de règles de décision

Lors de la phase de préparation des données, on associe à chaque pixel w de la base d'apprentissage, sa classe $C(w)$ qui peut être classé *peau* ou *non-peau*.

$$C : \Omega \rightarrow c = \{peau, non\_peau\}$$
$$w \rightarrow C(w)$$

La détermination de $C(w)$ n'est pas facile pour des raisons diverses liées aux conditions d'éclairage, aux différentes races etc. Nous cherchons un modèle de prédiction φ permettant d'identifier la classe C d'un pixel dont on ne connaît que les variables exogènes calculées dans la phase de préparation de données.

Pour déterminer l'espace de couleur optimal pour l'identification des pixels de peau, nous avons utilisé la méthode Relief [Kon94]. Cette dernière fournit des renseignements sur la pertinence d'une variable par rapport aux autres. Elle utilise les caractéristiques générales de l'ensemble d'apprentissage pour classer les variables en attribuant un poids, à chacune d'entre elles, compris entre -1 et 1.

3.1.3 Evaluation expérimentale

La construction de la base d'apprentissage est un élément important dans une démarche d'extraction de connaissances à partir des données. Nous avons travaillé sur

une base d'images composée du corpus CRL de Compaq (1ère base d'apprentissage disponible sur le sujet) et d'une base nommée ECL SCIV. Le corpus CRL [JR02a] est composé de 12 230 images conduisant à une base de 1.949.695.888 pixels. La base ECL SCIV [KTC03] est composée de plus de 1110 images de couleurs de peau issues de programmes TV conduisant à un modèle composé de 85 248 000 pixels.

Les expérimentations sur un échantillon d'instances composé de 3 412 992 pixels du corpus CRL, ont fait émergé la prédominance des composantes issues des deux espaces HSV, rgb (RGB normalisé) et du spectre (Dist) (voir tableaux 3.2 et 3.3) dans le processus de caractérisation de la peau.

Variables	H	Dist	r	S	b	Cb	I
Poids $\times 10^{-5}$	184	50	49	36	35	18	18

Variables	g	Cr	Q	R	Y	G	V	C	M	B
Poids $\times 10^{-5}$	-2	-11	-30	-48	-117	-119	-122	-134	-139	-155

FIGURE 3.2: Classement des variables par l'algorithme Relief.

Espaces de couleurs	HSV	rgb	Dist	YCbCr	YIQ	RGB	CMY
Poids $\times 10^{-4}$	$9,8$	$8,2$	5	$-1,10$	$-12,9$	$-32,2$	-39

FIGURE 3.3: Poids des espaces de couleur en utilisant l'algorithme Relief.

A partir de là, nous avons cherché à trouver les bonnes règles de décision à partir de ces espaces de couleur en utilisant les arbres de décision. En apprentissage automatique, la plupart des travaux s'appuient sur la théorie de l'information. Parmi les méthodes les plus populaires concernant les arbres de décision, on trouve la méthode ID3 de Quinlan [Qui86] et la méthode C4.5 [Qui93] qui est l'autre référence incontournable et la méthode SIPINA proposée par Zighed et Rakotomalala [ZR00].

Pour trouver le meilleur modèle de prédiction, nous avons testé ces techniques basées sur les graphes d'induction et une technique basée sur les réseaux de neurone. Pour cette dernière nous avons utilisé un perceptron à 2 couches cachées, composée chacune de 30 neurones, dont l'algorithme d'apprentissage est l'algorithme de propagation d'erreur classique. Les expérimentations ont été effectuées sur une base de test différente de celle qui a servi pour l'apprentissage, où nous avons calculé a chaque fois le taux de vrais positifs (VP), le taux de faux positifs (FP), et le taux d'erreur globale. Cette base de test est composée de 19 112 758 pixels extraits du corpus CRL de Compaq.

Nous avons effectué deux séries d'apprentissage, l'une en utilisant toutes les variables et l'autre uniquement avec les variables ayant un poids positif. Les tableaux 3.4 et 3.5 récapitulent les résultats obtenus sur chacune des deux séries avec les différents algorithmes.

Variables testées	Toutes les variables			
Algorithmes	C4.5	ID3	SIPINA	RN
Taux d'erreur Globale (TE)	18,77	18,39	17,78	21,17
Taux de vrais positifs	89,13	89,45	89,52	78,37
Taux de faux positifs	19 %	18,62	17,98	21,16

FIGURE 3.4: Résultats de l'utilisation de toutes les variables

Variables testées	H, Dist, r, S, b, Cb, I			
Algorithmes	C4.5	ID3	SIPINA	RN
TE Globale	18,51	18,45	17,33	20,47
Taux de vrais positifs	89,66	89,45	89,65	80,84
Taux de faux positifs	18,74	18,68	17,54	20,51

FIGURE 3.5: Résultats de l'utilisation des variables ayant un poids positifs

Les méthodes arborescentes ID3 et C4.5 procèdent par division et sont insensibles à la taille de l'échantillon. L'algorithme que nous avons adopté est proche de celui de SIPINA qui tente de réduire les inconvénients des méthodes arborescentes d'une part par l'introduction de l'opération de fusion et d'autre part par l'utilisation d'une mesure sensible aux effectifs. Celui-ci fournit une suite de partitions non nécessairement hiérarchisées.

Les résultats présentés précédemment nous ont conforté dans le choix de l'algorithme de SIPINA et nous ont permis de nous limiter aux 7 variables ayant un poids positif à savoir les composantes issues de HSV, rgb et Dist. On peut remarquer que cette réduction de variables a permis de diminuer légèrement le taux d'erreur.

Nous avons comparé nos approches bayesienne (A) et celle basée sur SIPINA avec les variables ayant un poids positif (B) et la méthode de classification proposée par M. Jones et J. Rehg pour Compaq (C) [JR02b]. Comme illustré dans la figure 3.6, notre approche basée sur les arbres de décision présente des améliorations par rapport à l'approche de référence utilisée par Compaq. Ces performances peuvent être encore plus significatives si l'on procède à une segmentation en zones dominantes.

Nous avons développé une solution baptisée **WebGuard**, à partir de ce modèle de peau, pour la classification et le filtrage de sites à caractère pornographique par un apprentissage qui s'appuie sur une combinaison judicieuse de plusieurs algorithmes de data mining avec non seulement une analyse du contenu textuel mais aussi du contenu structurel et visuel basé sur notre modèle de peau. Expérimenté sur une base de test de 400 sites composés de 200 sites adultes et 200 non adultes, WebGuard affiche un taux de classification de 97,4%. D'autres expériences sur une liste noire de 12 311 sites adultes, manuellement rassemblés et classifiés par le ministère de l'éducation français, montrent que WebGuard atteint un taux de classification de 95,62%.

3.1.4 Conclusion

Nous avons présenté dans cette section nos travaux pour l'extraction d'attributs pertinents de la couleur de peau. Ce travail a mis en évidence l'intérêt des outils de data

mining pour extraire les informations les plus pertinentes à partir de corpus de données diverses. Les résultats de nos expérimentations montrent que ce processus de fouille de données apporte un plus indéniable. L'étude comparative avec la méthode de référence utilisée par Compaq montre un gain d'efficacité substantiel, surtout lorsque la méthode est combinée avec une méthode de segmentation appropriée. Ce modèle de peau a été appliqué avec succès dans plusieurs applications liées à la détection de visages dans les séquences vidéo et dans le filtrage de sites.

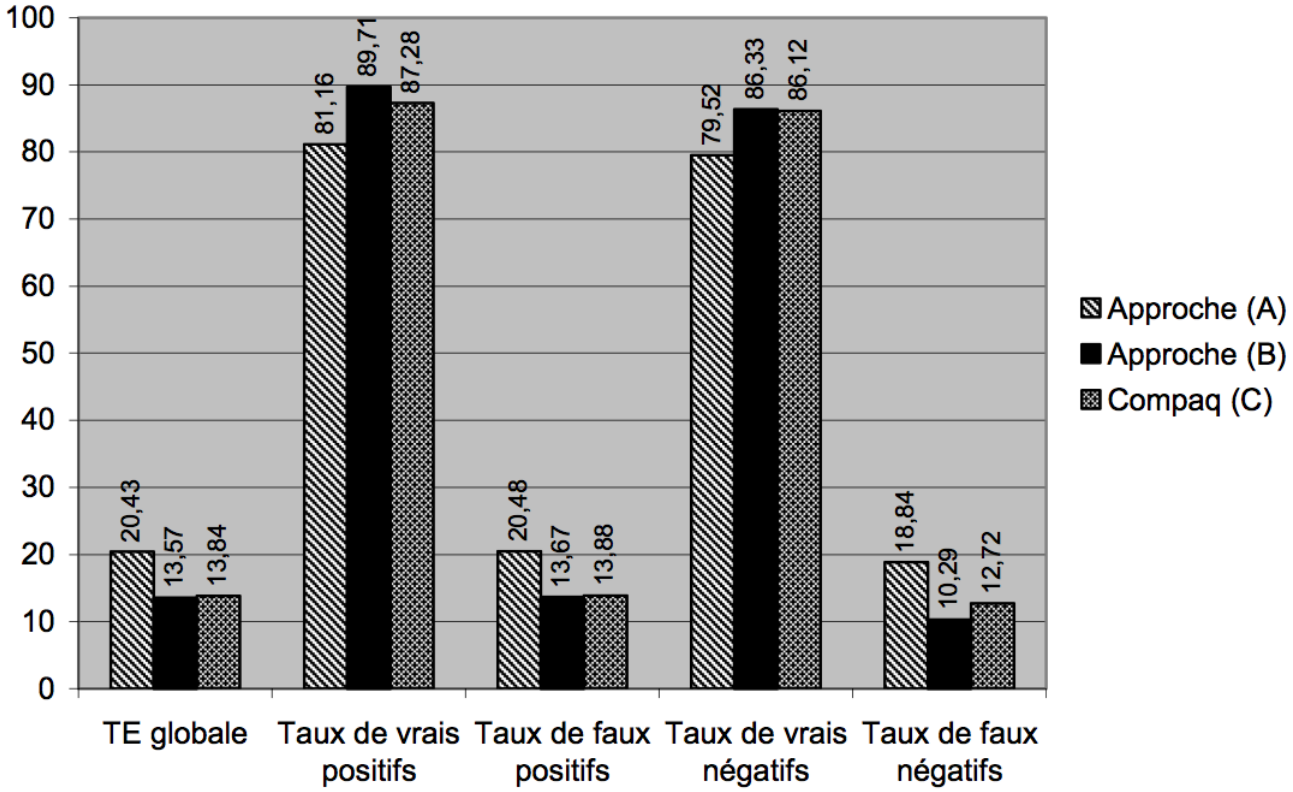


FIGURE 3.6: Comparaison de nos deux approches avec celle de M. Jones et J.Rehg (Compaq)

3.2 Segmentation par contours actifs

☞ Publications associées : [CL2, CI14, CI15, CI17, CN12, CN13, CO2] [ThèseF3] [StageR5] [StageR10]

3.2.1 Contexte et problématique

Nous nous intéressons à la recherche d'objets d'intérêt dans une séquence d'images (un visage, un objet en mouvement dans une vidéo par exemple). La segmentation qui nous intéresse ne peut pas être réalisée selon des critères bas niveau basés uniquement sur les pixels. Elle doit prendre en compte aussi bien les attributs perceptuels que d'autres contraintes contextuels ou géométriques. Dans ce contexte, les contours actifs apparaissent comme un outil approprié qui permet l'adjonction de contraintes et de connaissances a priori dans le processus de segmentation. Ils sont particulièrement bien adaptés pour la segmentation et le suivi d'objets en mouvement.La segmentation par contours actifs a l'avantage d'être un compromis entre régularité et adéquation aux données. De façon générale, l'énergie est composée de :

- Termes internes relatifs à la régularité géométrique du modèle. Ces termes sont indépendants de l'image et font intervenir des quantités différentielles (longueur, courbure, ...).
- Termes externes qui mettent en relation le modèle et l'image. Il s'agit de termes de contour, dépendant du gradient d'intensité [CKS97] intégrés le long de la courbe, ou de termes de région dépendant de statistiques globales calculées sur les domaines délimités par la courbe [CV01].

Les algorithmes de segmentation les plus robustes sont basés sur des fonctionnelles d'énergie et se distinguent par le type de la fonction d'énergie et par la technique d'optimisation utilisée pour la minimiser. Parmi les méthodes variationnelles, on peut citer les snakes [KWT88], les contours actifs géodésiques [CKS97], les régions actives géodésiques [PD05] et le modèle de Chan et Vese [CV01]. Ces premières méthodes procèdent via des descentes de gradients. Leur principal désavantage est la présence de minima locaux due à la non convexité des fonctionnelles d'énergies et la forte dépendance des résultats aux conditions initiales.

Afin de s'affranchir de ces limitations, il existe deux tendances de recherche. La première classe de travaux utilise différentes techniques d'optimisation (afin d'obtenir un minimum global de problèmes non convexes) tels que les graph-cuts [BFL06], [KT07], [KZ04]. Une alternative à ces travaux reformule l'énergie dans le but d'obtenir un problème convexe [BEV+07] [NEC06] [PCBC10] [LBS09] [PSG+08]. Ces techniques de segmentation sont basées sur une régularisation de type variation totale et essaient de trouver des fonctions caractéristiques qui minimisent la fonction de coût. Un minimum global est alors obtenu en effectuant simplement une descente de gradient. L'obtention d'un minimum global est important pour obtenir des algorithmes de segmentation robustes et indépendants de la position des contours initiaux.

Nous avons proposé une approche de contours actifs globaux inspirée de ces modèles pour la segmentation binaire en ne considérant qu'un seul objet d'intérêt.

3.2.2 Modèle convexe binaire

Les méthodes orientées objet supposent que l'image f est composée d'une région de fond Ω_{ext} et d'une région disjointe Ω_{int} représentant l'objet (voir figure 3.7).

Une formulation générique des modèles des contours actifs basés régions [JBBA01a] peut être adoptée dans le cadre des contours actifs multiples et de la classification d'image. La fonctionnelle d'énergie à minimiser s'écrit alors :

$$J(\Gamma, \alpha_1, \alpha_2) = \mu \int_{\Gamma} g(\Gamma(s))ds + \lambda \left[\int_{\Omega_{int}} r^{\alpha_1}(x)dx + \int_{\Omega_{ext}} r^{\alpha_2}(x)dx \right] \tag{3.3}$$

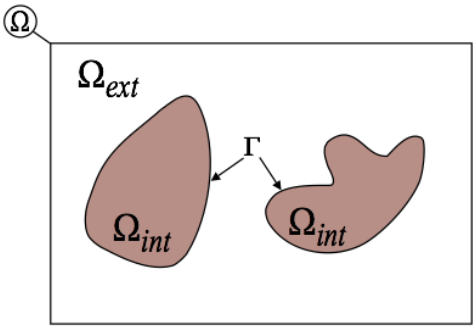


FIGURE 3.7: Domaine image, Ω subdivisée en une région de fond,Ω_{ext} et une région correspondant à l'objet à reconstruire Ω_{int}. Le contour actif Ω, marque la frontière entre les deux régions.

Les fonctions $r^{\alpha} : \Omega \rightarrow \mathbb{R}$ désignent les descripteurs géométriques de chaque région qui dépendent des paramètres de régions $\alpha = (\alpha_1, \alpha_2))$ tels que des scalaires [CV01], des vecteurs [PD02] [ZY96] ou des fonctions [TJW01][VCTC02].

Cas 1 : Quand $\lambda = 0$, on retrouve le contour actif géodésique.

Cas 2 : Pour $g = Id$ $(g(x) = 1))$ et $r^{\alpha_i}(x) = (I_{(x)} - s_i)^2 + |\nabla(s_i)|$ on retrouve le modèle de Mumford-Shah [TJW01] [VCTC02] [MS89], où s_1 et s_2 sont deux fonctions qui approximent l'image à l'intérieur et à l'extérieur des régions.

Cas 3 : Pour $g = Id$ et $r^{\alpha_i}(x) = (I_{(x)} - c_i)^2$, on retrouve le modèle de Chan et Vese [CV01] qui consiste à minimiser la fonctionnelle suivante : $J(\Gamma, \alpha_1, \alpha_2) = \mu \int_{\Gamma} ds + \lambda \left[\int_{\Omega_{int}} (I_{(x)} - c_1)^2 dx + \int_{\Omega_{ext}} (I_{(x)} - c_2)^2 dx \right]$ où c_1, c_2 représentent les moyennes de l'image I à l'intérieur et à l'extérieur de la courbe.

Cas 4 : Pour $r^{\alpha_i} = logP_i(I \backslash \alpha_i)$ on retrouve le modèle de Paragios et Deriche [PD02].

Tous les modèles présentés et d'autres qui leur sont similaires de part leur construction sont fortement non convexes, et les méthodes de descente de gradient pour la minimisation d'énergie qui aboutissent aux équations d'évolution ne garantissent pas la convergence vers un minimum global. Elles sont donc très sensibles à l'initialisation. Pour pallier à ces problèmes, de nouvelles méthodes de contours actifs ont été récemment proposées [She06][NEC06][BEV+07].

L'approche de segmentation que nous proposons repose sur une minimisation globale et sur la formule de la co-aire [Cha05] pour rendre convexe ce problème initialement non convexe et trouver un minimum global.

Une région est représentée implicitement par une fonction indicatrice :

$$u : V \rightarrow \{0,1\}$$
$$x \rightarrow \begin{cases} 1 & si \quad x \in \Omega_{int} \\ 0 & sinon \end{cases}$$

$V = \Omega_{ext} \cup \Omega_{int}$.

Soit $u = 1_{\Omega_{int}}$ la fonction caractéristique de la région Ω_{int}.

$1 - u = 1_{\Omega_{ext}}$ est la fonction caractéristique de la région Ω_{ext}.

La convexification consiste à prendre le problème relaxé qui s'écrit sous la forme suivante :

$$\underset{u(x)\in[0,1]}{Min} \left\{ \mu \int_{\Gamma} g(x)|\nabla u(x)|dx + \int_{\Omega} u(x) r^{\alpha_1}(x)dx + \int_{\Omega} (1-u(x)) r^{\alpha_2}(x)dx \right\} \tag{3.4}$$
$$= \underset{u(x)\in[0,1]}{Min} \left\{ \mu \int_{\Gamma} g(x)|\nabla u(x)|dx + \lambda \int_{\Omega} u(x) \left[r^{\alpha_1}(x) - r^{\alpha_2}(x) \right] dx \right\}$$

L'énergie 3.3 peut être globalement optimisée et sa minimisation permet d'extraire l'objet et le fond. La figure 3.8 (resp. figure 3.9) montre quelques résultats illustrant le cas 3 du modèle de Chan et Vese (resp. le cas 4 du modèle de Paragios et Deriche) avec une initialisation arbitraire par une ellipse.

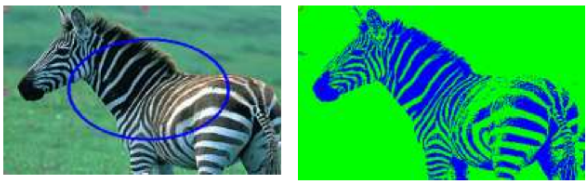


FIGURE 3.8: Approche de Chan-Vese : Segmentation selon la moyenne.

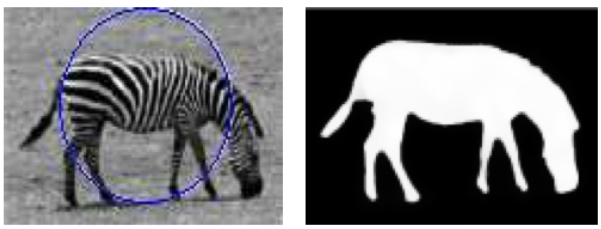


FIGURE 3.9: Approche de Paragios-Deriche : Segmentation selon la densité de probabilité.

Nous avons proposé une segmentation binaire en minimisant l'énergie (3.4) avec $r^{\alpha_i}(x) = logP_i(x)$ où P_i est la fonction de densité de probabilité de la région Ω_i qui est définie par :

$$P_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{(I(x)-\mu_i)^2}{2\sigma_i^2}}. \tag{3.5}$$

Comme on peut constater dans l'image 3.11, le modèle basé sur la densité de probabilité est intéressant par sa robustesse au bruit grâce notamment à la prise en compte des statistiques du bruit et d'informations à priori (homogénéité, textures). C'est le modèle que nous avons adopté pour la détection d'objets visuels dans une vidéo.

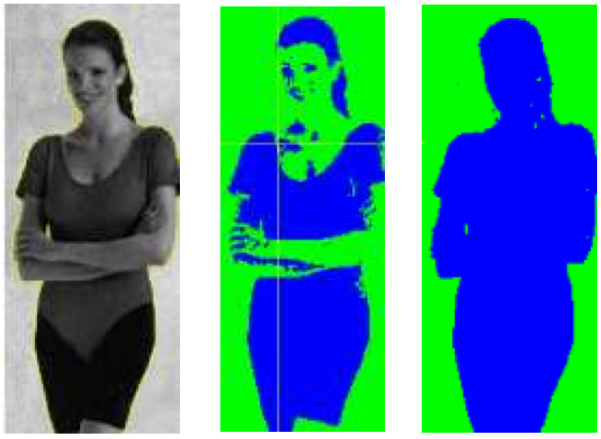


FIGURE 3.10: Segmentation de l'image (à gauche) en utilisant la moyenne (milieu) et avec utilisation de la probabilité de densité (à droite).

La figure 3.9 montre la segmentation d'une image en utilisant une initialisation avec des contours multiples en utilisant le modèle basé sur la densité de probabilité.

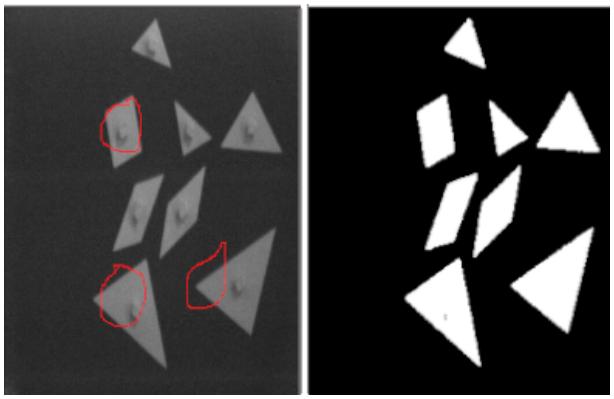


FIGURE 3.11: Segmentation de l'image Tangram en utilisant une initialisation multiple.

La méthode converge plus vite quand les contours initiaux sont proches de l'objet d'intérêt, comme l'illustre le tableau 3.1 correspondant à l'image 3.12. Le contour initial donne une première idée sur *l'objet* et *le fond*. La segmentation est accélérée quand l'un des contours initiaux est proche de l'objet d'intérêt.

FIGURE 3.12: Segmentation de l'image du Cameraman [256 × 256] par différentes initialisations.

Contour Initial	Nombre d'itérations
C1	5
C2	4
C3	6
C4	3

TABLE 3.1: Convergence de la segmentation (en nb d'itérations) de la figure 3.12.

On peut utiliser une forme a priori comme un terme de fidélité additionnelle, qui rend la segmentation plus robuste aux occlusions, et aux données hétérogènes. De façon générale, la contrainte de forme est introduite par une métrique permettant de comparer le contour actif à l'instant t avec la forme à priori.

La segmentation d'objets en mouvement est une étape primordiale avant toute analyse d'une scène. Une revue des méthodes de détection de mouvement peut être trouvée dans [ZL01].

3.2.3 Segmentation d'un objet en mouvement

Un objet d'intérêt n'est pas toujours très visible et très distinguable du fond. Toutes les méthodes de segmentation trouvent leurs limites quand il y a une occlusion ou quand l'objet possède les mêmes caractéristiques statistiques tels que la moyenne et la texture que son voisinage ou le fond.
Notre contribution en segmentation des objets en mouvement dans une vidéo, réside dans la segmentation en volume par contour actif en tenant compte d'un critère supplémentaire issu du calcul du flot optique.
Une façon d'estimer le mouvement dans une séquence d'images est de calculer le flot optique. Soit l'équation de contraintes du flot optique : $\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0$.
Résoudre cette équation est un problème mal posé. Pour effectuer ce calcul, on utilise en général un terme de régularisation sur le flot, par exemple un a priori est que le flot est lisse comme dans les travaux fondateurs d'Horn et Schunck [HS81] ou de Lucas et Kanade

[LK81]. Nous cherchons donc le flot optique (u, v) qui minimise la fonctionnelle suivante :

$$E(u,v) = \int_{\Omega} \left(\left(\frac{\partial I}{\partial x} u + \frac{\partial I}{\partial y} v + \frac{\partial I}{\partial t} \right)^2 + \alpha \psi \left(|\nabla u|^2 + |\nabla v|^2 \right) \right) dxdy \tag{3.6}$$

où ψ est une fonction croissante différentiable, $\nabla = (\partial x, \partial y)^T$ l'opérateur gradient et α est un paramètre de régularisation. Le choix de la fonction ψ influence le processus de régularisation et donc les résultats de l'estimation de mouvement.

Les méthodes d'estimation et de segmentation conjointes du mouvement définissent un critère dépendant à la fois de la région et du mouvement à estimer. La minimisation de ce critère peut être traitée de différentes façons : comme une minimisation hiérarchique non linéaire [MP02], comme un problème aux valeurs propres ou par l'évolution d'un contour actif [PD00].

Dans ce qui suit, nous présentons notre approche de segmentation d'un objet en mouvement. Un calcul du flot optique par la méthode de Horn & Schunk est préalablement effectué. La norme du flot calculé est ensuite incorporée dans le modèle de contour actif global basé sur la minimisation d'une fonctionnelle d'énergie qui prend aussi en compte l'information photométrique. La phase d'estimation de mouvement par flot optique permet d'avoir une idée sur les zones où il y a une activité.

Les fonctions r^{α_1} et r^{α_2} de l'équation 3.4 décrivent l'a priori sur les statistiques des données sur Ω_{int} et Ω_{ext}. Pour la détection d'un objet en mouvement, l'importance se situe au niveau du terme décrivant la région extérieure au contour. Nous avons traité plusieurs façons de combiner les contours actifs et le flot optique :

Cas 1 : Détecteur de bords dans une zone d'activité : ($g = \frac{1}{1+x^2}, r^{\alpha_1} = \beta$ et $r^{\alpha_2} = \|F\|$). On retrouve le modèle proposé par F.Ranchin et F. Dibos [RD04]. La différence est que l'on ne considère pas une image de fond et que l'on introduit le détecteur de bords $g(\nabla I)$. L'idée est donc de contrôler la norme du flot sur la région extérieure, tout en vérifiant la régularité de la frontière et en empêchant celle-ci de traverser des zones homogènes de l'image . $\mu g(\|\nabla\|)$ vise à régulariser la frontière tout en l'attirant vers un bord.

Cas 2 : Estimation du fond de la séquence d'images : (g=Id, $r^{\alpha_1} = \beta$ et $r^{\alpha_2} = |B-I|$). On retrouve le modèle proposé dans Jehan Besson, Barlaud et Aubert [JBBA01b]. Ce descripteur est basé sur la différence entre une image et le fond estimé à partir de plusieurs images de la séquence (B). Nous avons proposé une variante (g=Id, $r^{\alpha_1} = \beta$ et $r^{\alpha_2} = \|F\|$) qui utilise la norme du vecteur du flot optique de la séquence.

La constante β représente un seuil de référence de l'amplitude du flot au-delà de laquelle on considère qu'il y a mouvement. Les figures suivantes 3.13 montrent le résultat de la segmentation vidéo obtenue par la combinaison du contour actif et Horn & Schunk dans le cas 1 du paragraphe 3.2.3 sur un échantillon de corpus connus.

Nous montrons également l'efficacité de cette approche sur un corpus de vidéos d'activités humaines (courir, sauter,...) (figure 3.14) qui sera présentée dans le paragraphe 3.3.

FIGURE 3.13: Segmentation vidéo avec estimation du fond $\beta = 40$ et $\lambda = 10$.

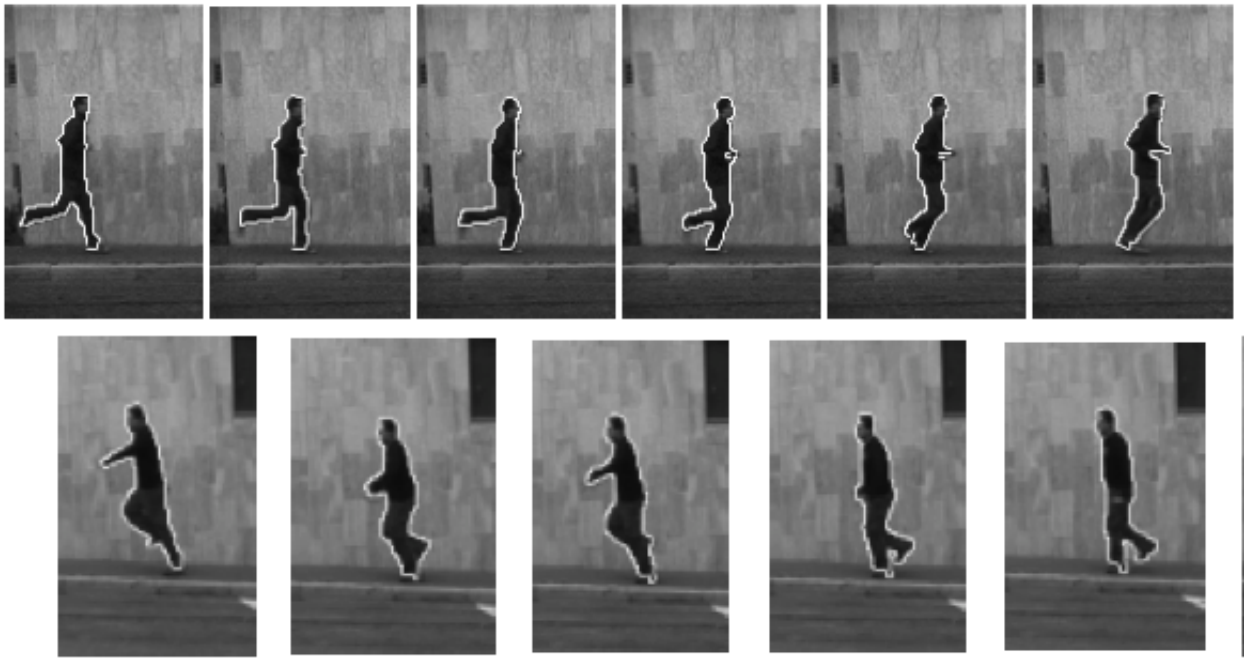


FIGURE 3.14: Segmentation de personnes en action avec $\beta = 120$ et $\lambda = 15$.

3.2.4 Conclusion

Dans cette section, nous avons présenté notre méthode des contours actifs basée sur un modèle convexe binaire pour la segmentation d'un objet d'intérêt (volume) dans une séquence vidéo. Nous avons combiné cette méthode avec le flot optique pour proposer un algorithme de segmentation 2D + t d'un objet d'intérêt en mouvement. La formulation énergétique du modèle de contour actif a été étendue par l'intégration d'un critère supplémentaire issu du calcul du flot optique. L'approche présentée est moins sensible aux contours initiaux et au bruit. L'exploitation seule des contours actifs peut être réalisée sur chaque image 2D et le contour final trouvé peut être utilisé comme un contour initial

de l'image suivante. Mais, les résultats peuvent être spatialement incohérents. Nous avons préféré travailler directement sur la séquence entière en tenant compte des contraintes temporelles liées au mouvement. L'évaluation a montré que l'approche proposée donne de bons résultats, en terme de temps de calcul et d'extraction d'un objet d'intérêt complexe. Néanmoins, plusieurs critères caractérisant les propriétés des objets d'intérêt restent à prendre en compte, tels que les descripteurs de forme, de texture ou même le type de mouvement. L'utilisation d'une énergie de la forme à priori en 3D par template de la forme [FZ05] ou des poses [BKT06] reste possible.

3.3 Analyse de gestes

☞ Publications associées : [RI2, CI1, CI2, CI9, CI13, CN1, CN4, CO2, , CN3, CO3][StageR4] [StageR6]

3.3.1 Contexte et problématique

La problématique de l'analyse des gestes et des actions a émergé à l'issu des mes collaborations avec des partenaires en sciences cognitives et en neuro-sciences en particulier avec F Jouen [1] et M. Molina [2] du pôle Modesco (Modélisation en sciences cognitives) à Caen.

Le premier travail en commun a consisté à interpréter un geste du toucher en fonction des propriétés des objets afin d'adapter les traitements et développer des habiletés motrices chez certains malades atteints d'infirmité motrice cérébrale (IMC). Le geste est décomposé en procédures exploratoires élémentaires : Mouvement de la main - Pression sur l'objet. Ensuite, nous nous sommes intéressés à la classification de actions de personnes. Les objectifs recherchés consistent à classer les gestes filmés à partir de descripteurs pertinents du comportement observé et à évaluer l'impact de la régularisation discrète sur graphe (réduction de la dimension et débruitage) sur le processus de classification et donc sur la qualité des résultats obtenus.

Les formes d'un objet d'intérêt sont couramment représentées par une surface, un contour ou un volume. Les méthodes de représentation de formes spatio-temporels décrites dans la littérature peuvent être répertoriées en trois classes :

Représentation volumique structurelle : L'objectif de ces méthodes est de caractériser un objet 3D comme une suite d'objets élémentaires. Partant des observations de Biederman sur la reconnaissance humaine par composants [Bie87], Bergevin et Levine [BL93] définissent 36 formes de base ou "géons" (géométrique ions).

Représentation statistique de la forme qui présente des variabilités dans l'espace et dans le temps. Pour tenir compte de cette variabilité, il est possible de décrire la forme de référence de façon statistique à partir de plusieurs modèles d'un même objet. Ces approches consistent à caractériser les objets 3D par plusieurs distributions de descripteurs de formes qui peuvent être locaux telles que la courbure en tous points de l'objet [ZP03] [ABP03], ou globaux comme les histogrammes ou les invariants [TV08].

Représentation globale de la forme : Dans cette classe, on distingue deux types d'approches : Les transformées qui visent à déterminer des représentations définies en terme de transformation intégrale telles que les transformées de Hough 3D [KPVG10], ou de Radon [Fid85] et les approches basées sur les moments qui peuvent être définies comme la projection de la fonction définissant l'objet sur un ensemble de fonctions caractéristiques. Ces approches ont été utilisées avec succès en reconnaissance de forme 2D. On peut citer les moments géométriques, Legendre, Fourier-Mellin, Zernike, ART. Plusieurs de ces moments ont été étendus en 3D : Fourier 3D [OK06], Ondelette 3D [PR99] et Zernike 3D [Can99].

1. Equipe CHART : Cognitions Humaine & ARTificielle) de l'EPHE à Paris.
2. Equipe PALM : Psychologie des Actions Langagières et Motrices.

Nous considérons dans cette partie des modélisations globales, d'un geste ou d'une action, basées sur les moments spatio-temporels et qui sont exprimées sous formes d'intégrales sur l'intérieur du domaine de l'objet d'intérêt délimité par sa silhouette.

3.3.2 Moments spatio-temporels

Nous avons proposé plusieurs contributions basées sur l'utilisation de différents modèles des moments (les moments géométriques et les moments de Krawtchouck) pour caractériser un volume binaire. Dans cette partie, nous ne présentons que les moments spatio-temporels de Zernike que nous avons évalué sur différentes bases avec et sans la méthode de régularisation (avec $p = 2$) présentée dans le chapitre précédent.

L'idée de base est de représenter l'ensemble des vidéos par un graphe $G = (V, E, w)$ où $V = \{v_1, v_2, \ldots, v_n\}$ est l'ensemble des vidéos (volumes) $(2d+t)$. Chacune est caractérisée par une fonction caractéristique $f(v_i)$. A partir du volume binaire d'une vidéo v, on extrait des caractéristiques représentatives de l'action basée sur les moments spatiotemporels de Zernike. Ces moments sont considérés parmi les moments invariants les plus efficaces en terme de performance globale pour la représentation et reconstruction des formes. Shutter et Nixon [SN01] ont montré que ces moments présentent un bon taux de reconnaissance et une description compacte pour caractériser une forme en mouvement. Les moments de Zernike 3D ont l'avantage de capturer l'information globale de la forme 3D sans exigence de la fermeture des contours [NK03].

Les fonctions de Zernike 3D, initialement défini par Canterakis [Can99], sont définies par :

$$Z^v_{nlm}(x, y, t) = R_{nl}(r) \cdot Y_{lm}(\theta, \phi) \tag{3.7}$$

où $R_{nl}(r)$ est le terme radial, et $Y_{lm}(\theta, \phi)$ est le terme angulaire. Z^v_{nlm} peut être ré-écrite sous une forme compacte comme une combinaison linéaire des moments géométriques d'ordre n

$$Z^v_{nlm}(x, y, t) = \sum_{p+q+r \leqslant n} \mathcal{X}^{pqr}_{nlm} x^p y^q t^r \tag{3.8}$$

Soit $g(x, y, t)$ l'ensemble des points appartenant au volume associé à v où x, y et t représentent les coordonnées spatio-temporelles. Z^v_{nlm} forme un système orthonormal complet. Il est possible d'approximer la fonction d'origine g par un nombre fini de moments de Zernike 3D Ω^v_{nlm} comme suit :

$$g(x, y, t) = \sum_{n=0}^{\infty} \sum_{l=0}^{n} \sum_{m=-l}^{l} \Omega^v_{nlm} Z^v_{nlm}(x, y, t) \tag{3.9}$$

Les moments de Zernike 3D sont définis par :

$$\Omega^v_{nlm} = \frac{3}{4\pi} \sum_{p+q+r \leqslant n} (-1)^m \mathcal{X}^{pqr}_{nlm} m^v_{pqr} \tag{3.10}$$

où pour $k = (n-1)/2$:

$$\begin{aligned}\mathcal{X}_{nlm}^{pqr} =& c_{lm} 2^{-m} \sum_{s=0}^{k} q_{kls} \sum_{\alpha=0}^{s} \binom{s}{\alpha} \sum_{\beta=0}^{s-\alpha} \binom{s-\alpha}{\beta} \sum_{r=0}^{m} (-1)^{m-r} \binom{m}{r} \\ & i^r \sum_{\mu=0}^{(l-m)/2} (-1)^{\mu} 2^{-2\mu} \binom{l}{\mu} \binom{l-\mu}{m+\mu} \sum_{s=0}^{\mu} \binom{\mu}{s},\end{aligned} \tag{3.11}$$

et le facteur de normalisation :

$$c_{lm} = \frac{\sqrt{(2l+1)(l+m)!(l-m)!}}{l!}, \tag{3.12}$$

et

$$q_{kls} = \frac{(-1)^k}{2^{2k}} \sqrt{\frac{2l+4k+3}{3}} \binom{2k}{k} (-1)^s \frac{\binom{k}{k} \binom{2(k+l+s)+1)}{2k}}{\binom{k+l+s}{k}} \tag{3.13}$$

m_{pqr}^v sont des moments géométriques d'ordre $(p+q+r)$ du volume binaire représentés par :

$$m_{pqr}^v = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \sum_{t=0}^{N_t-1} x^p y^q t^r g(x,y,t) \tag{3.14}$$

m_{000}^v représente le volume de l'objet et $(m_{100}^v, m_{010}^v, m_{001}^v)$ sont les coordonnées du centre de l'objet.
Le choix de l'ordre maximal des moments de Zernike 3D est crucial. Les moments d'ordres supérieurs offrent une caractérisation fine pour une forme donnée, mais peuvent devenir plus sensibles au bruit. Le descripteur de forme $f(v)$ de la vidéo (v) est la norme du vecteur Ω_{nlm}, défini par les $(2l+1)$ moments selon la formule suivante :

$$f(v) = \{\mathcal{V}_{nl}^v = \|\Omega_{nlm}^v\| \ : \ n \in [0,N], l \in [0,n], m \in [-l,l]\} \tag{3.15}$$

La distance entre deux vidéos u et v est calculée via leurs moments de Zernike 3D :

$$\|f(u) - f(v)\| = \|\mathcal{V}_{nl}^u - \mathcal{V}_{nl}^v\| = \sqrt{\sum_{n=0}^{N} \sum_{l=0}^{n} (\mathcal{V}_{nl}^u - \mathcal{V}_{nl}^v)^2} \tag{3.16}$$

[RLB09] ont proposé une mesure de similarité robuste basée sur les moments de Zernike qui prend en compte l'information de phase des moments complexes et qui renvoie un angle optimal de rotation entre les deux images.

3.3.3 Validation expérimentale

3.3.3.1 Exploration tactile

∗ **Contexte et Objectif :**

L'identification d'un geste manuel est une tâche difficile du fait de la double fonction des mains. Les mains ont, en effet, deux fonctions imbriquées : une fonction motrice, dans

laquelle les perceptions tactiles aident à la réussite des actions, et une fonction perceptive, dans laquelle la motricité est au service de la perception pour identifier et reconnaître les objets [RDLTH01] [Gru08].

Selon Klatzky et Lederman [LK87], l'identification et la reconnaissance des propriétés des objets sont obtenues par l'exécution de procédures exploratoires élémentaires (PEs). Selon ces auteurs, l'exploration tactile est une activité séquentielle réalisée via l'exécution de mouvements stéréotypés des doigts et de la paume des mains. Ces mouvements sont intentionnels et dépendent de la propriété d'objet que le système tactile choisit de traiter :
• Frottement latéral de la surface pour la texture.
• Pression pour la consistance.
• Enveloppement et suivi des contours pour la forme et la taille.

Deux propriétés différentes (texture et consistance) ont été testées. La texture de l'objet pourrait être lisse ou granuleuse. La consistance pourrait être dure ou molle. Le mouvement latéral sur un objet lisse correspond à un frottement de la main le long de sa surface. Nous avons défini quatre procédures exploratoires élémentaires :

- Mouvement Latéral pour Objets Lisse (LMSO)
- Mouvement Latéral pour Objets Rugueux (LMGO)
- Pression sur Objet Doux (PSO)
- Pression sur Objet Dur (PHO)

La figure 3.15 présente un exemple de la base d'objets manipulés et de certains gestes de toucher.

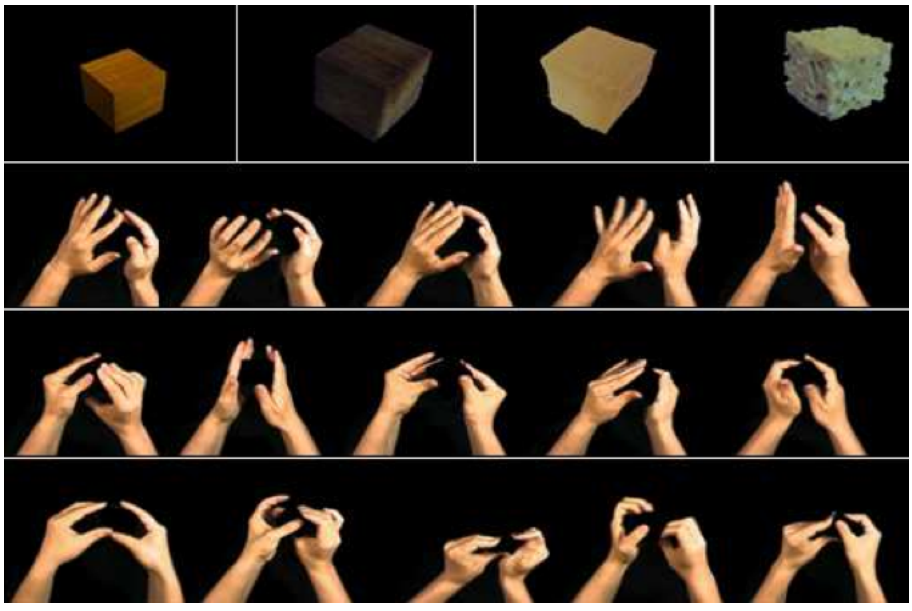


FIGURE 3.15: Objets manipulés et illustration de quelques gestes de toucher.

Les gestes de la main sont capturés par une caméra web.

Nous avons utilisé une base d'apprentissage composé de 120 vidéos qui représentent les quatre actions (PEs) : 20 vidéos de PSO, 29 vidéos de LMGO, 37 vidéos de PHO et

34 vidéos de LMSO. Nous avons testé notre approche sur une base de test composée de 34 nouvelles vidéos : 7 PSO, 7 LMGO, 10 PHO et 10 vidéos LMSO. Chaque vidéo est composée de 111 frames.

L'objectif de notre contribution est de développer une méthode d'analyse vidéo capable d'interpréter des gestes de toucher et de classer les objets selon leur consistance et leur texture. Nous avons utilisé une méthode de réseaux de neurones par rétro-propagation composé de trois couches : la première couche se compose de 14 nœuds qui prend comme données les moments spatio-temporels, la couche cachée contient 30 neurones et la dernière est constituée de deux neurones pour reconnaitre l'une des 4 PEs étudiées. Dans [RI1] nous avons également évalué l'apport de la régularisation des données pour l'amélioration des résultats.

∗ **Segmentation et classification :**
D'abord, nous avons exploité la couleur de la peau pour la segmentation des régions d'intérêt (ROI) en utilisant des règles de décision citées dans la section 3.1. Pour la segmentation, nous avons utilisé une méthode de segmentation semi-supervisée pour segmenter les images vidéo en trois parties : main gauche, main droite et l'arrière-plan.

Pour chaque vidéo u, nous calculons $f(u)^{left}$ (resp. $f(u)^{right}$) les moments spatio-temporels de Zernike correspondants à la main gauche (resp. main droite). Ainsi, la comparaison entre deux vidéos u et v caractérisées par leurs moments spatiotemporels de Zernike $f(u)$ et $f(v)$ représentant deux gestes quelconques est définie par :

$$\frac{\|f(v)^{right} - f(u)^{right}\| + \|f(v)^{left} - f(u)^{left}\|}{2} \tag{3.17}$$

Un premier résultat de classification consiste à reconnaître une main gauche d'une main droite en utilisant les moments spatiotemporels de Zernike. Le taux de reconnaissance atteint 100%. La figure 3.16 illustre bien cette séparation des volumes des deux mains. Cette figure visualise l'organisation d'un échantillon de la base des mains gauches et droites selon le vecteur de Fiedler.

Pour la reconnaissance des PEs, nous avons utilisé des réseaux de neurones par rétro-propagation. Sur les 34 vidéos test, le réseau de neurones reconnaissait bien 82.3% des gestes dans les vidéos. Nous avons proposé récemment [RI1] une amélioration de ces résultats en utilisant la réduction de la dimension par marches aléatoires et la régularisation de ces données dans l'espace réduit composé des trois premiers vecteurs propres (cf. section 2.4). Avec cette régularisation, le taux de reconnaissance atteint 87%.

Le résultat de reconnaissance des quatre PEs est résumé dans le tableau 3.2.

	Taux de reconnaissance (%)	
Procédure exploratoire	sans régularisation	avec régularisation
LMSO	90	93.6
LMGO	87.5	89.2
PSO	71.4	77.4
PHO	80	85.9

TABLE 3.2: Reconnaissance de gestes en utilisant les moments spatio-temporels de Zernike avec et sans la régularisation

FIGURE 3.16: Réorganisation des mains selon le vecteur de Fiedler après régularisation spectrale.

3.3.3.2 Actions

On peut trouver dans [AR11] une synthèse des différents travaux sur la reconnaissance des actions proposées dans la littérature. Plusieurs travaux ont été effectués pour classer les actions dans les vidéos en utilisant différentes méthodes telles que les réseaux de neurones [BMW+10], GMM [CTL+10] etc.

Les expérimentations que nous avons menées ont été effectuées sur deux corpus connus que sont les bases Weizmann et KTH. La 1^{ere} base Weizmann proposée par Blanck et al. [GBS+05] est généralement utilisée pour tester des modèles de classification d'actions. Elle contient 81 vidéos avec une basse résolution (180×144). Ces vidéos contiennent 9 personnes qui effectuent 9 actions différentes. Chaque séquence vidéo représente une action unique. Des exemples pour chacune de ces actions sont présentés dans 3.17. Le calcul des moments spatio-temporels nécessite un prétraitement des vidéos qui consiste à segmenter la vidéo en un volume spatio-temporel et extraire les silhouettes associées.

Nous avons choisi d'utiliser une variante de SVM [SV99] (LS-SVM) pour classer les actions en se basant sur les moments spatio-temporels de Zernike. Différents ordres de moment ont été testés afin de trouver un ordre optimal qui est **l'ordre 7**. Les résultats de classification sont présentés dans le tableau 3.18.

FIGURE 3.17: Exemples d'actions de la base Weizmann (à gauche) et de la base KTH (à droite).

	walk	run	skip	jack	jump1	jump2	wave1	wave2	bend
walk	1.0	.00	.00	.00	.00	.00	.00	.00	.00
run	.02	.98	.00	.00	.00	.00	.00	.00	.00
skip	.00	.03	.97	.00	.00	.00	.00	.00	.00
jack	.00	.00	.00	1.0	.00	.00	.00	.00	.00
jump	.10	.00	.00	.00	.83	.07	.00	.00	.00
jump in place	.00	.00	.02	.00	.00	.98	.00	.00	.00
wave with one hand	.00	.00	.00	.00	.02	.00	.96	.00	.02
wave with two hands	.00	.00	.00	.00	.00	.00	.04	.96	.00
bend	.00	.00	.00	.00	.00	.00	.01	.00	.99

FIGURE 3.18: Matrice de confusion matrix de la base Weizmann.

Méthode	Taux de reconnaissance
Zelnik et Irani 2001[ZmI01]	66.66 %
Vezzani et al. 2010[VBC10]	86.7 %
Dhillon et al. 2009 [DNL09]	88.55 %
Ta et al. 2010 [TWL+10a]	94,5 %
notre approche	96.33 %
Kellokumpu et al. [KZP11]	98.9 %

FIGURE 3.19: Comparaison avec d'autres approches sur la base Weizmann

Nous avons également mené des expériences sur la base de données d'actions KTH. Cette base contient six types d'actions humaines, chacune effectuée à plusieurs reprises par 25 personnes. Elle présente plus de complexité que celle de Weizmann à cause des grandes variations des angles de vue, des échelles et des apparences.

Sur cette base, le processus de régularisation était bénéfique. Nous procédons à une réduction de la dimension à partir de la matrice des transitions P calculée à partir des descripteurs basés sur les moments spatio-temporels de Zernike. Nous projetons les vidéos sur l'espace réduit construit à partir des premiers vecteurs porteurs d'informations. Ensuite, nous construisons un graphe de k-ppv dans l'espace réduit qui sera sujet de débruitage par régularisation. L'idée est de procéder à un lissage des données en tenant compte de la proximité spatiale afin de faciliter le regroupement de noeuds similaires.

Soit $G_r = (V_r, E_r, w_r)$ le graphe construit dans l'espace réduit. V_r est l'ensemble des noeuds représentant les vidéos. w_r est la fonction de poids qui reflète la proximité entre les sommets dans l'espace réduit calculée à partir de leurs coordonnées propres.

Chaque sommet v est défini par un triplet (x_v, y_v, z_v) qui indique la position du point dans l'espace $\{\lambda_2\psi_2\}$, $\{\lambda_3\psi_3\}$ et $\{\lambda_4\psi_4\}$.

$f(v)$ est une fonction qui représente la proximité spatiale : f(v) = v $=(x_v, y_v, z_v)$ et $\|f(u) - f(v)\| = \|u - v\|$.

La figure 3.20 présente la projection de la variété de la base KTH dans l'espace tridimensionnel réduit, tandis que la figure 3.21 montre le résultat du lissage des coordonnées de ces données.

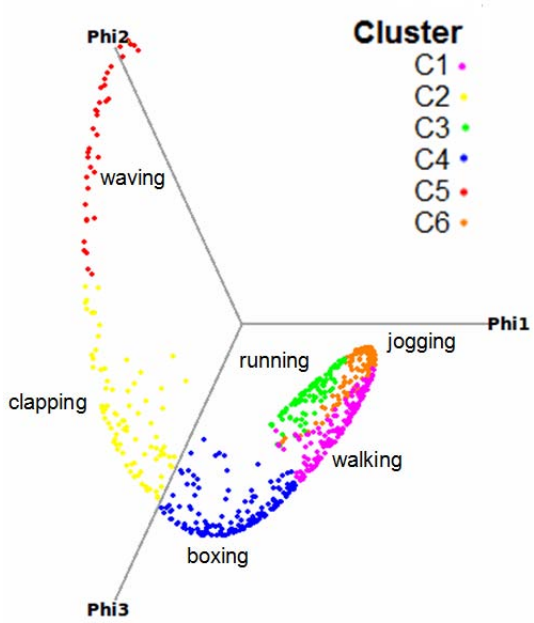


FIGURE 3.20: Projection de la base KTH dans l'espace réduit

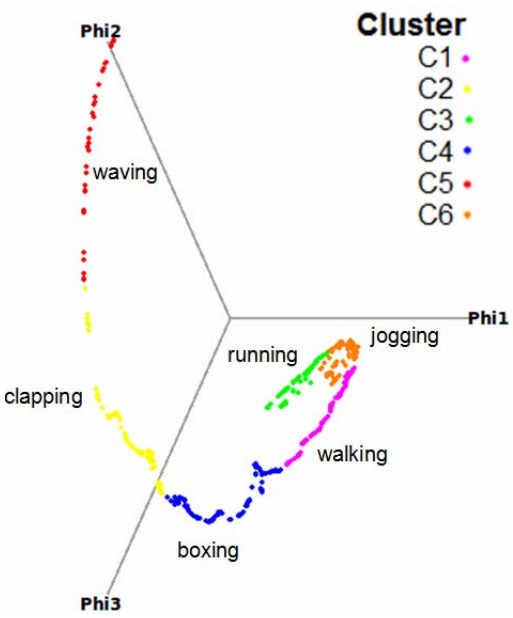


FIGURE 3.21: Projection de la base KTH dans l'espace réduit régularisé

Le tableau 3.22 affiche la matrice de confusion obtenue pour la base KTH. Nous pouvons observer que les résultats obtenus avec la régularisation sont bien meilleurs que sans la régularisation.

walking	.94/**1**	.01	.04	.00	.00	.00
running	.11	.89/**.98**	.00	.00	.00	.00
jogging	.01	.08	.87/**.89**	.00	.00	.04
handclapping	.00	.00	.00	.92/**.96**	.04	.00
handwaving	.14	.00	.00	.02	.83/**.94**	.01
boxing	.00	.00	.00	.03	.02	.95/**.98**
	walking	**running**	**jogging**	**handclapping**	**handwaving**	**boxing**

FIGURE 3.22: Matrice de confusion de la base KTH.

La comparaison avec d'autres méthodes récentes de la littérature confirme l'intérêt de notre approche et le bon taux de reconnaissance qui atteint une moyenne de 95,17% en utilisant la régularisation des données dans l'espace réduit de faible dimension.

Méthode	Taux de reconnaissance
Notre approche (avec régularisation)	95.17 %
Kim et al. 2007[kKfWC07]	95.16 %
Xinghua et al. 2009 [XMH09]	94.0 %
Kellokumpu et al. 2011 [KZM11]	93.77 %
Ta et al. 2010 .[TWL+10a]	93.0 %
Ning et al 2008 [NHW+08]	92.31 %
Ballan et al. 2009 [BBDB+09]	92.17 %
Wong et al. 2007 [WKC07]	91.6 %
Costantini et al. 2011 [CSS+11]	91.17 %
Notre approche (sans régularisation)	90.7 %
Dhillon et al. 2009 [DNL09]	84.67 %
Nienles et al. 2006 [NWFf06]	81.50 %
Dollar et al. 2005 [DRCB05]	81.17 %

FIGURE 3.23: Comparaison avec d'autres approches sur la base KTH

Les résultats expérimentaux montrent un bon taux de classification pour la plupart des actions. Néanmoins, le choix de l'ordre optimal pour le moment de Zernike reste un problème ouvert.

3.3.4 Conclusion

Nous avons présenté un processus de représentation et de reconnaissance des données liées aux gestes et aux actions humaines. Nous avons résumé chaque vidéo par un volume binaire que nous avons décrit par les moments spatio-temporels de Zernike. La méthode que nous avons présenté est basée sur les graphes et est composée de plusieurs étapes :

- Extraction du volume binaire de chaque vidéo.
- Représentation (Reconstruction) par les moments spatio-temporels de Zernike.
- Construction du graphe $G = (V, E, w)$ de la base dans l'espace initial. La similarité $w(i, j)$ entre deux vidéos i et j est calculée à partir de leurs moments spatio-temporels.
- Réduction de la dimension et projection dans l'espace réduit $\{\lambda_2\psi_2\}$, $\{\lambda_3\psi_3\}$ et $\{\lambda_4\psi_4\}$.
- Construction du graphe $G_r = (V_r, E_r, w_r)$ de la base dans l'espace réduit. La similarité $w_r(i, j)$ entre deux vidéos est calculée à partir de leurs positions spatiales dans l'espace réduit.
- Lissage du graphe G_r (nuage des points).
- Classification des noeuds ou partitionnement du graphe en classes.

Nous avons cherché à interpréter dans un premier temps des gestes de toucher en vidéo et en déduire des propriétés de texture et de consistance des objets. Un processus de reconnaissance est présenté en utilisant les marches aléatoires et les réseaux de neurones. Ensuite, nous avons montré nos résultats de classification des actions sur les bases de données Weizmann et KTH avec le classificateur LS-SVM.

3.4 Analyse du regard

☞ Mes publications associées : [CI3, CI7, CN7] [ThèseF2]

3.4.1 Contexte et problématique

L'analyse du regard sous l'angle de la psychologie expérimentale (l'occulométrie cognitive) s'intéresse à la fois au geste moteur lui-même (les mouvements des yeux), à la façon dont on explore les scènes visuelle par le regard ainsi qu'à l'influence du regard sur le comportement d'autrui. Estimer ou prédire la zone de l'écran observée par un spectateur durant son expérience multimédia présente un intérêt applicatif important tant en termes de traitement de l'image (reframing, compression adaptative ...) qu'en termes de marketing (pertinence d'une publicité, ...) ou d'interfaces homme-machine (IHM) (défilement automatique d'un contenu par exemple, sélection par le regard).

Nous avons abordé cette problématique dans le cadre du projet "l'œil et la main" (une partie du projet ENEIDE [1]) où l'objectif consiste à doter l'enseignant d'un outil d'évaluation de l'attention des élèves qui permet d'analyser les fixations et les saccades oculaires lors de la présentation de documents.

Le problème du suivi du regard a été étudié et développé depuis de nombreuses décennies en raison de ses usages potentiels dans de nombreuses applications. Au début, les méthodes de suivi du regard qui étaient particulièrement intrusives limitaient les mouvements spontanés de l'utilisateur : une bobine magnétique intégrée dans une lentille posée sur l'œil [YS75], des électrodes placées autour de l'œil pour mesurer l'activité musculaire [Sha67], des lunettes avec capteur infrarouge [Duc07]. Aujourd'hui le système intrusif le plus répandu comprend un casque contenant deux micro caméras placées devant l'œil de l'utilisateur : l'une pour capter l'image de l'écran d'ordinateur et l'autre pour capter l'image de l'œil [LWP05]. Il existe également d'autres systèmes qu'un semi réfléchissant placé devant l'écran couplé à un système de rotation permettant de suivre les mouvements de l'utilisateur, de manière à garder une image fixe reflétée par le miroir dans une caméra. Les difficultés de mise en œuvre des méthodes intrusives les réservent à des application de suivi du regard dans les seuls laboratoires de recherche. Elles ne peuvent donc pas convenir à des applications grand public.

Il existe des systèmes non-intrusifs [SYW97] [BP94] [PCG+03] [HNL06] [MKAF99] [YC05] [OMY02] [DCB] qui utilisent uniquement une caméra avec des méthodes de traitement d'images et des réseaux de neurones pour la détection et l'évaluation de la direction du regard. Ces systèmes sont sensibles au mouvement de la tête et nécessitent une phase de calibration. Pour résoudre le problème du mouvement de la tête, il existe des systèmes [ZJ05] [OM04] [OMK03] qui utilisent deux caméras pour faire la calibration et un infrarouge pour obtenir la position des yeux par réflexion. Aussi, il est apparu une gamme de produits commerciaux non intrusifs tels que *MyTobii*, *Visioboard*, *EyeGaze* qui proposent des mesures très précises avec une liberté de mouvement mais dont le prix reste très élevé (entre 15.000 et 30.000 euros [MB09]).

1. Espace Numérique Educatif Interactif de Demain

L'objectif affiché du travail de Ba Linh NGUYEN est de développer un système de suivi du regard qui pallie aux inconvénients listés précédemment.

3.4.2 Descripteurs de Haar avec AdaBoost

Pour suivre le regard d'un utilisateur, nous devons d'abord détecter la position de ses yeux. La stratégie utilisée se fonde sur une cascade de classifieurs de type *AdaBoost* pour sélectionner un nombre restreint de bons *descripteurs de Haart* comme dans l'approche initialement proposée par Paul Viola [VJ01] et améliorée par Rainer Lienhart [LM02]. Les classifieurs sont formés avec des milliers d'images positives et négatives basées sur des descripteurs de Haar. La cascade de classifieurs est construite pour accroître la performance de la détection et réduire les temps de calcul.

Le principe consiste à combiner un ensemble de fonctions de classification faibles pour former une fonction de classification plus efficace. La fonction de classification est faible si elle est seulement capable de reconnaître deux classes au moins aussi bien que le hasard ne le ferait. Après la première phase d'apprentissage, les exemples sont recombinés de sorte que celles qui ont été mal classifiées par la précédente fonction de classification faible auront le poids le plus grand. La fonction de classification forte finale prend la forme d'un perceptron.

Considérons le problème d'apprentissage, dans lequel un grand ensemble de fonctions de classification est combiné en utilisant un vote majoritaire pondéré. Le défi est d'associer un grand poids à chaque bonne fonction de classification et un plus petit poids aux fonctions faibles.

FIGURE 3.24: Un descripteur sélectionné par AdaBoost.

Nous avons utilisé la méthode de détection de Harris pour détecter un ensemble de points caractéristiques des composantes faciales tels que les coins (des yeux, du nez et de la bouche). On a fait le choix de ne garder que les extrémités externes de ces composantes. la figure 3.25 montre ces extrémités que l'on suit à travers les frames par la méthode basée sur le flot optique de suivi de Lucas et Kanade.

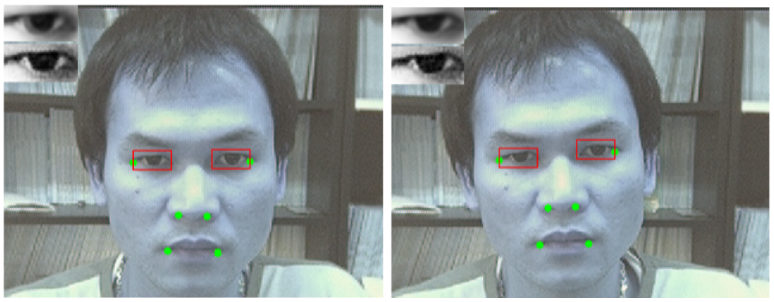


FIGURE 3.25: Résultat de la détection des coins sur les yeux, le nez et la bouche.

Lors du suivi en temps réel, certains points disparaissent et d'autres apparaissent. Nous avons proposé d'utiliser un algorithme d'analyse de la régularité des données basé sur le principe des **données aberrantes** en se basant sur la proximité des points.

Une donnée est aberrante quand elle s'écarte trop du comportement régulier des autres données. La définition des outliers par leur distance est une approche populaire pour trouver des exemples inhabituels dans un ensemble de données [AP02] [RRS00]. La méthode que nous avons utilisé consiste à prendre les K voisins les plus proches d'une mesure. Si le point n'est pas dans le voisinage, la mesure est considérée comme inhabituelle. Pour récupérer les points perdus, nous avons choisi de favoriser l'exemple qui a le maximum de voisins.

3.4.3 Prédiction du regard par processus Gaussien

Nous présentons dans cette section comment localiser le regard de l'utilisateur sur l'écran. Nous disposons d'une base de données (issues de la phase de calibration) qui comprend en **entrée** x_i l'image de l'œil, de taille 32×16 et en **sortie** y_i la position du point du regard sur l'écran.

Nous appelons cette base de données $\mathcal{D} = \{(x_i, y_i)|i = 1, \cdots, n\}$ où x_i désigne le vecteur d'entrée (co-variables) et y_i désigne une cible (variable dépendante), et n est le nombre d'observations.

Soit X la matrice regroupant toutes les co-variables x_i et y le vecteur des cibles. On peut écrire $\mathcal{D} = (X, y)$. Nous nous sommes intéressés aux inférences que l'on peut faire sur les relations entre les entrées et les cibles (cf. figure 3.26).

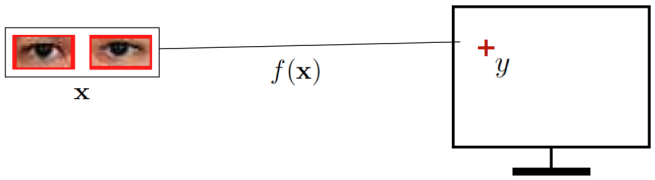


FIGURE 3.26: Relation entre les yeux de l'utilisateur et la position de son regard sur l'écran.

Une grande variété de méthodes a été proposée en littérature pour résoudre ce problème. Ces méthodes utilisent le plus souvent un grand ensemble de données d'apprentissage et des réseaux de neurones pour entraîner la fonction. Comme il est difficile d'avoir une base d'apprentissage avec toutes les entrées possibles, nous avons décidé d'utiliser un processus gaussien pour calculer la distribution prédictive pour $f_* \equiv f(x_*)$ avec $x_* : p(f_*|x_*, \mathcal{D})$.

Un Processus Gaussien (PG) est entièrement défini par sa fonction moyenne et sa fonction de covariance. Nous définissons la fonction de moyenne $m(\mathbf{x})$ et la fonction de covariance $k(\mathbf{x}, \mathbf{x}')$ d'un processus réel $f(\mathbf{x})$ comme

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], \end{aligned} \tag{3.18}$$

et le processus gaussien s'écrit alors :

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big). \tag{3.19}$$

Pour simplifier la notation nous considérons la fonction moyenne égale à zéro.

Dans notre cas, les variables aléatoires représentent la valeur de la fonction $f(\mathbf{x})$ à la position $\mathbf{x}$. Souvent, les processus Gaussiens sont définis sur le temps. Ce n'est pas le cas dans notre utilisation du PG. Par souci de simplification, nous utilisons la notation $f_i \triangleq f(\mathbf{x}_i)$ pour identifier la variable aléatoire qui correspond au cas $(\mathbf{x}_i, y_i)$.

Un exemple simple d'un processus Gaussien peut être obtenu à partir du modèle Bayésien de régression linéaire $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$ avec la distribution a priori $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$. Nous avons la moyenne et la covariance :

$$\begin{aligned} \mathbb{E}[f(\mathbf{x})] &= \phi(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = 0, \\ \mathbb{E}[f(\mathbf{x}) f(\mathbf{x}')] &= \phi(\mathbf{x}^\top) \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \phi(\mathbf{x}') = \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}'). \end{aligned} \tag{3.20}$$

$f(\mathbf{x})$ et $f(\mathbf{x}')$ sont des gaussiennes jointes avec une moyenne de zéro et une covariance donnée par $\phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}')$. En effet, les valeurs des fonctions $f(\mathbf{x}_1), ..., f(\mathbf{x}_n)$ correspondant à un nombre de points d'entrée quelconque n sont joint Gaussien.La fonction de covariance indique la covariance entre les paires de variables aléatoires [Ras06].

$$\operatorname{cov}\big(f(\mathbf{x}_p), f(\mathbf{x}_q)\big) = k(\mathbf{x}_p, \mathbf{x}_q) = \exp\big(-\frac{1}{2\ell^2}|\mathbf{x}_p - \mathbf{x}_q|^2\big). \tag{3.21}$$

où ℓ est *caractéristique de l'échelle des longueurs.*

La spécification de la fonction de covariance implique une distribution sur des fonctions. En pratique, on choisit un certain nombre de points d'entrée X_* et on génère un vecteur aléatoire gaussien avec cette matrice de covariance :

$$\mathbf{f}_* \sim \mathcal{N}\big(\mathbf{0}, K(X_*, X_*)\big) \tag{3.22}$$

Dans le cas des observations sans bruit, la distribution conjointe des sorties de formation $\mathbf{f}$ et des sorties de test $\mathbf{f}_*$ selon l'a priori est :

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right). \tag{3.23}$$

Pour obtenir la distribution a posteriori sur les fonctions, nous avons besoin de limiter a priori cette distribution conjointe pour ne conserver que les fonctions qui sont en accord avec les points observés. Ainsi :

$$\begin{aligned} \mathbf{f}_* | X_*, X, \mathbf{f} \ \sim \mathcal{N}\big(&K(X_*, X) K(X, X)^{-1} \mathbf{f}, \\ &K(X_*, X_*) - K(X_*, X) K(X, X)^{-1} K(X, X_*)\big). \end{aligned} \tag{3.24}$$

Dans le cas sans contraintes, on suppose que les observations $f(x)$ sont bruitées par un bruit additif ϵ ($y = f(x) + \epsilon$). Nous supposons que ce bruit possède une distribution indépendante de moyenne zéro et de variance σ_n^2 : $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$

L'a priori sur le bruit observé devient :

$$\text{cov}(y_p, y_q) = k(\mathbf{x}_p, \mathbf{y}_q) + \sigma_n^2 \delta_{pq} \quad \text{ou} \quad \text{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 I, \tag{3.25}$$

où δ_{pq} est un delta de Kronecker qui est égal à 1 si $p = q$ et 0 autrement.

En introduisant le terme de bruit dans l'éq. (3.23), on peut écrire la distribution conjointe des valeurs cibles observées et des valeurs de la fonction de test sous l'a priori de la manière suivante :

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right). \tag{3.26}$$

La prédiction du regard pour la nouvelle entrée $\mathbf{x}_*$ peut être obtenue par :

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)\big) \tag{3.27}$$

où

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y}, \tag{3.28}$$

$$\text{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*). \tag{3.29}$$

3.4.4 Validation expérimentale

Avant de faire une session de suivi du regard, l'utilisateur effectue une procédure de calibration pour créer une base d'apprentissage.

Une série de tests (avec $\ell = 500$) ont été conduites sous deux conditions différentes :

– Tête de l'utilisateur est stable.

– Situation normale avec mouvement de tête.

En situation normale, nous effectuons une dizaine de procédures de calibrations sous différents angles de vue comme l'illustre la figure 3.27.

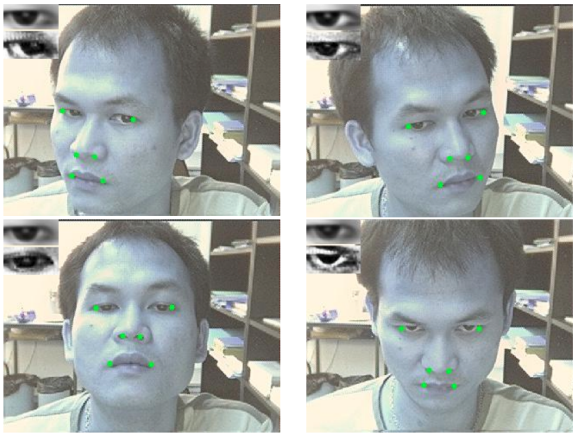


FIGURE 3.27: Calibration sous différents angles de vue.

La figure 3.28 présente les résultats de prédiction du regard dans deux situations. La première correspond au mouvement des yeux avec une position stable et la seconde en

situation libre de mouvement. Les points triangulaires sont les points cibles du test, les autres points sont les points prédis par un processus gaussien. Les résultats obtenus sont très encourageants.

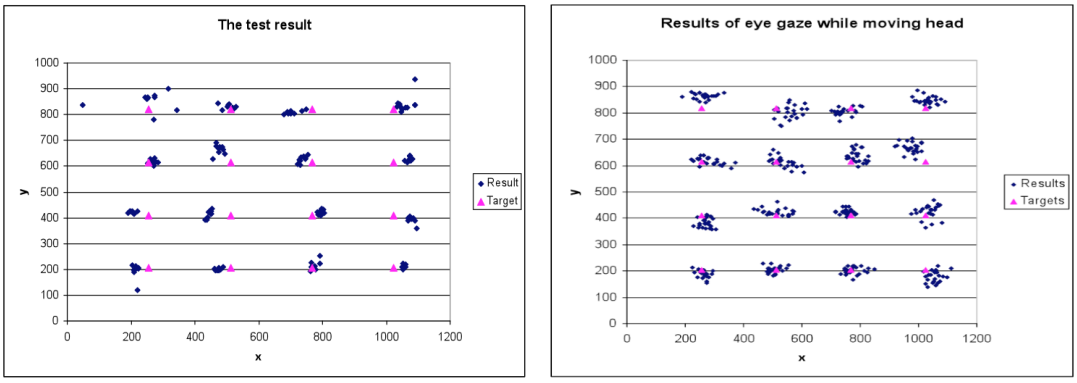


FIGURE 3.28: De gauche à droite : Résultat avec tête stable et situation libre.

3.4.5 Conclusion

Nous avons présenté une nouvelle approche de capture et de suivi du regard en temps réel dans le cadre de la communication homme machine. L'idée principale est de combiner la méthode de boosting AdaBoost avec l'algorithme de Lucas-Kanade pour le calcul du flot optique et les processus gaussiens.

Les systèmes actuels ont besoin de créer une base d'apprentissage assez grande (environs 10 000 paires de données). Plusieurs systèmes utilisent deux caméras pour estimer la distance avec l'utilisateur et distinguer le mouvement de la tête et des yeux. Le résultat actuel de la capture du regard dépend d'une vérité terrain obtenue par une calibration qui n'est pas très contraignante. Avec cette nouvelle approche, nous avons pu résoudre plusieurs problèmes inhérents aux systèmes de capture et de mesures oculométriques en temps réel.

4

Conclusion

Sommaire

Ce mémoire retrace mon activité professionnelle depuis la soutenance de ma thèse. Ce manuscrit m'a permis de faire un bilan de mes recherches et je présente dans cette section une synthèse des contributions ainsi que les perspectives de mes travaux.

4.1 Synthèse des contributions

Les contributions que nous avons proposées et décrites dans ce mémoire sont les suivantes :

Chapitre 2 : *Régularisation sur graphe - Restauration et caractérisation de séquences d'images et des vidéos.*

Dans ce chapitre, j'ai abordé en première partie la régularisation non-locale des images et des vidéos et ses applications au débruitage et à la simplification. En seconde partie, j'ai présenté la diffusion spectrale et montré une autre manière de définir un ordre de données par une réduction non linéaire de dimension grâce à des marches aléatoires sur graphe.

* Partie I : Restauration des données
  - Nous avons proposé l'extension d'un cadre de régularisation discrète aux séquences d'images et de vidéos et nous avons proposé des algorithmes de régularisation non locale sur graphes par p-Laplacien, avec $1 \leqslant p \leqslant \infty$, pour le débruitage et la simplification des séquences d'images.
  - Nous avons proposé deux solutions pour le problème de la restauration des données manquantes (inpainting). Tout d'abord, la restauration par régularisation. Il s'agit d'une extension du modèle présenté en débruitage. La seconde approche est basée sur les fonctions p-harmoniques sur graphe.
* Partie II : Régularisation spectrale par marches aléatoires sur graphe
  - Nous avons proposé un algorithme de segmentation par une réduction non linéaire de dimension à partir des points caractéristiques de l'image.
  - Nous avons proposé une méthode de caractérisation des éléments saillants dans une image.

– Nous avons proposé une formulation du mécanisme du bouclage de pertinence basée sur la diffusion spectrale pour affiner les résultats de la recherche dans une base d'algues.

Chapitre 3 : *Analyse Vidéo - Segmentation et reconnaissance de gestes.*

Dans ce chapitre, j'ai abordé en première partie le développement d'outils qui permettent de caractériser au mieux les pixels de peau, de segmenter et de suivre un geste ou un individu en mouvement. En seconde partie, j'ai abordé l'analyse de gestes et du regard d'une personne. J'ai présenté une approche de classification de bases vidéo à partir des descripteurs spatio-temporels par SVM et une méthode de prédiction du regard par processus gaussiens.

* Partie I : Caractérisation et Segmentation
  – Nous avons proposé un modèle prédictif basé sur les arbres de décision pour l'extraction des attributs pertinents concernant la couleur de la peau.
  – Nous avons proposé des méthodes de segmentation d'un objet en mouvement par contours actifs. Une première contribution concerne la segmentation en volume en utilisant une segmentation binaire convexe. La seconde contribution concerne la segmentation en prenant en compte les contours actifs et le flot optique.
* Partie II : Classification et Interprétation
  – Nous avons proposé une approche de reconnaissance de gestes de toucher (base propriétaire de l'équipe PALM) et des actions (base Weizmann et KTH) à partir des moments de Zernike 3D. Une contribution récente a été présentée concernant l'apport de la régularisation sur graphe construit dans l'espace réduit de faible dimension.
  – Nous avons présenté une méthode de capture du regard basée sur Adaboost et les descripteurs de Haar. L'approche utilise la méthode de Lucas et Kanade pour suivre les points caractéristiques du visage. Par ailleurs, nous avons proposé une méthode de prédiction du regard par processus gaussiens. La distribution est estimée à partir d'un ensemble de tests occulométriques lors de la calibration.

4.2 Perspectives

Il y a deux nécessités dans le traitement d'images et de vidéos : d'abord une formalisation rigoureuse, et la résolution de vrais problèmes. La deuxième motive la première, qui, à son tour, justifie la seconde. En continuité avec les travaux que j'ai déjà réalisés, mon projet de recherche s'oriente d'abord vers l'analyse de séquences d'images liées aux activités humaines telles que des gestes et actions.

4.2.1 Reconnaissance des actions

La reconnaissance de l'activité humaine est un champ de recherches très actif et particulièrement complexe dans le domaine de la vision par ordinateur. Les applications sont nombreuses notamment pour la vidéo-surveillance, l'interaction homme-machine, la recherche et l'archivage de vidéos, le diagnostic médical ou encore l'analyse sportive. La reconnaissance d'actions nécessite l'extraction d'attributs pertinents parfois multi sources qui sont généralement dépendants de l'application. Quelle que soit leur provenance, les attributs sont généralement entachés d'imprécisions dues aux capteurs (e.g. la caméra) ou aux algorithmes de traitement des données. Plusieurs verrous scientifiques restent à lever :

A la recherche de nouveaux descripteurs : La majorité des méthodes récentes repose sur l'extraction d'un ensemble de descripteurs (essentiellement locaux tels que des descripteurs SIFT) [CCA09], la séquence (plan-séquence) étant alors représentée par un sac contenant ces descripteurs. Nous avons proposé d'utiliser des descripteurs globaux basés sur les moments spatio-temporels de Zernike. Il existe d"autres types de moments tels que les moments orthogonaux discrets de Tchebichef et de Krawtchouck qui constituent des pistes à explorer.

Les aspects de saillance et d'impacts émotionnels sont très souvent ignorés dans les approches récentes et pourtant peuvent aider à catégoriser les images et les vidéos. Ces deux aspects que j'ai étudiés à partir des expressions faciales dans les images peuvent être étendus à la vidéo afin d'intégrer une autre expression du comportement dans la caractérisation des actions.

Fusion de descripteurs : La fusion d'informations est un domaine qui connaît une évolution importante, en particulier avec la multiplication des sources d'informations qu'il s'agisse de capteurs, d'informations à priori ou de descripteurs, etc. Deux tendances sont amenées à co-exister :

Combinaison de classifieurs : Cette voie peut être envisagée pour fiabiliser la reconnaissance en exploitant la complémentarité qui peut exister entre les classifieurs. Sur ce point, la littérature abonde de travaux présentant des méthodes de combinaison qui se différencient aussi bien par le type d'informations apportées par chaque classifieur que par leurs capacités d'apprentissage et d'adaptation [PAF+09].

Fusion de descripteurs : Prendre en compte tous les descripteurs, les considérant comme compétitifs et complémentaires, pour construire un nouveau vecteur. Ce dernier peut être construit en utilisant la théorie des probabilités [MYC+10], ou par réduction de dimensionnalité. De nombreuses méthodes de réduction de dimension non-linéaire ont été proposées [LV07][TSL00] qui exploitent généralement un graphe de voisinage défini dans

l'espace initial de grande dimension. L'ensemble de ces méthodes exploite une analyse spectrale du Laplacien sur graphe et souffre par conséquent du difficile passage à grande échelle car l'analyse spectrale est basée sur un calcul matriciel. Le p-Laplacien est une généralisation non-linéaire du Laplacien sur graphe. Il reçoit l'attention de la communauté en apprentissage à travers la relation existant entre le p-Laplacien sur graphe et les coupes de Cheeger [BH09]. Récemment, Luo et al [LHDN10] ont proposé une méthode permettant une décomposition complète en vecteurs propres du p-Laplacien. Une des perspectives naturelles qui se dessine alors est l'étude des potentialités d'une analyse spectrale du p-Laplacien. Nous avons utilisé une formulation discrète du p-Laplacien dans le domaine spatio-temporel. Il reste à étudier le lien entre les deux formulations dans les domaines spectral et spatial.

Relations entre descripteurs : Il existe peu de descripteurs en vidéos qui bénéficient conjointement de l'information spatiale et temporelle. Deux approches de comparaison des descripteurs sont proposées dans la littérature : La première méthode dite de bag-of-features (BoF) consiste à quantifier l'espace des descripteurs de manière à produire un dictionnaire de mots visuels. Chaque descripteur du sac de la séquence est alors projeté sur le(s) mot(s) le(s) plus proche(s) afin de constituer une représentation vectorielle (vecteur de fréquence d'apparition des mots visuels), facilement exploitable par les outils de classification [KMLS10] [vdSGS10][WXDL11]. La seconde consiste à établir directement une mesure de similarité entre les sacs de descripteurs sous la forme de noyaux [GHS11]. Cette méthode a l'avantage de fournir directement une mesure exploitable par les outils d'apprentissage statistique largement utilisés aujourd'hui, tels que les SVM.

Avec ces descripteurs, la structure des actions n'est pas prise en compte et peu de travaux traitent des relations spatiales/ temporelles entre les descripteurs [TWL+10b]. Certaines approches proposent d'intégrer a priori dans la description une information sur les relations spatiales [SAK11][HGBRM10].

Le formalisme basé sur les graphes de topologie arbitraire pourrait s'avérer pertinent pour pallier à l'imprécision de la représentation par description locale. En effet, la position et le nombre des points d'intérêt sont très sensibles au bruit ou au filtrage rendant l'existence d'une relation spatiale avec ses voisins non certaine d'une séquence à l'autre. Aussi, certaines relations spatiales entre points peuvent être perturbées par la transformation qui existe entre deux images, comme le changement de point de vue par exemple. Si l'on considère une collection d'images à laquelle nous associons un graphe de topologie arbitraire dont les noeuds sont décrits par des descripteurs de dimension quelconque, nous pouvons alors considérer une régularisation et/ou interpolation non locale directement pour ces données [ZWBL11] [Luo06]. En plus, on peut intégrer des relations temporelles telle que celles d'Allen dans la similarité entre les noeuds. [UPL10][XHZX12]. Il faudrait dans ce cadre étudier l'influence de la topologie des graphes sur le temps de calcul [LL06].

4.2.2 Recherche par le contenu

Avec le développement des grands réseaux de contenu, il y a un changement majeur dans la façon d'envisager l'accès à l'information. Les domaines traditionnels se sont développés autour du concept d'un utilisateur isolé, de l'idée d'un corpus de documents multimédias indépendants les uns des autres. C'est l'utilisateur qui fait la connexion entre

les unités d'information lors de sa recherche. Ces données multimédias sont par essence hétérogènes même si leur contenu prépondérant est visuel. Elles sont organisées en réseaux avec des relations multiples entre les éléments. Les relations peuvent être explicites (e.g. amis) ou implicites et liées à des proximités sémantiques entre utilisateurs et / ou éléments (partage d'intérêts thématiques, comportements similaires, etc). Ces réseaux peuvent être modélisés par des graphes de contenus, avec des relations multiples (hypergraphes) et des nœuds représentant différents types de contenus (graphes hétérogènes). Les usages et les méthodes liés pour l'accès à l'information doivent être reconsidérés à la lumière de ces nouveaux paradigmes. L'apprentissage automatique est appelé à jouer un rôle majeur pour l'analyse des traces d'usage et entre les éléments de contenu (des individus, des images, des objets, etc).

Les techniques actuelles d'indexation et de recherche de séquences d'images par le contenu visent à extraire automatiquement des caractéristiques visuelles des images et à les organiser dans des index multidimensionnels pour ensuite faciliter la recherche dans les grandes bases d'images. Les techniques d'indexation ont pour but d'organiser l'ensemble des descripteurs généralement hétérogènes et de grande dimension afin que les procédures de recherche soient performantes. Cette organisation se traduit généralement par une structuration des descripteurs. Les techniques de recherche par le contenu quant à elles, consistent à développer et à appliquer des outils qui permettent de sélectionner les séquences les plus pertinentes par leurs contenus, en terme de similarité.

Les principaux défis ouverts sont :

Catégorisation en unités homogènes : Les différents problèmes de recherche et de catégorisation des informations peuvent se formaliser comme le calcul de scores associés aux graphes. Plusieurs familles de méthodes issues des graphes peuvent être employées comme les marches aléatoires, les méthodes par régularisation ou l'utilisation de noyaux de diffusion sur les graphes. Cette tâche n'est pas une tâche triviale puisque l'on mélange des informations sémantiques, structurelles et visuelles.

Formalisation de nouveaux opérateurs adaptés au traitement de données complexes. La plupart des techniques sur graphe, comme celles que nous avons proposées dans ce manuscrit, ont été développées dans le cas de graphes simples (un seul type de relation) pour des graphes homogènes (toutes les entités sont de même type) et doivent donc être adaptées au traitement de données plus complexes.

Apprentissage semi-supervisé : révèle à partir d'instances multiples, d'exemples multi-facettes ou avec des étiquettes multiples ou bruitées, de nouveaux problèmes d'apprentissage. Une direction qui nous intéresse correspond à des applications comme l'annotation où il faut associer à chaque élément du graphe des étiquettes multiples avec éventuellement un ordre sur les étiquettes. La seconde direction correspond au bouclage de pertinence qui consiste à ordonner les nœuds du graphe eux-mêmes en fonction de la requête de l'utilisateur.

Dans cette thématique, le processus de diffusion par marches aléatoires paresseuses (il est possible de rester dans un même état à chaque étape de la marche) me semble une approche plus adaptée à l'hétérogénéité des noeuds ce qui relie fortement cette approche à celles utilisant les temps de commutation.

Au niveau applications on s'intéressera essentiellement à la participation aux challenges dédiés à l'analyse de comportements et à la structuration d'activités humaines

en vidéo. Un des objectifs est d'interpréter les comportements individuels (actions observables). Comme la tâche de reconnaissance de comportements requiert une analyse sémantique, parfois complexe, de ce qui apparaît dans l'image, elle représente le défi le plus coriace pour les systèmes d'analyse vidéo. L'analyse et l'interprétation de comportements nécessitent de reconnaître des patrons de mouvements et d'en dégager, à un plus haut niveau, une description des actions et interactions.

4.2.3 Applications

Mes collaborations pluridisciplinaires tissées avec les partenaires d'autres équipes et du monde professionnel continueront à être fructifiées et renforcées en particulier sur l'analyse du regard.

De nouvelles tendances se dessinent en analyse du regard. On propose d'évaluer les potentialités d'un signal physiologique Electro-OculoGramme (EOG) pour une estimation alternative du trajet du regard sur un écran. Potentiellement moins contraignante qu'un Eye Tracker, cette méthode exploiterait les variations de champs électriques induites par le mouvement de l'œil et mesurées dans l'EOG pour estimer la position ou le mouvement courant du regard. Par ailleurs, l'automatisation de tests symptomatiques comme les tests du nystagmus, de convergence oculaire et de réaction à la lumière des pupilles permet d'évaluer les facultés d'une personne.

Dans ce contexte, on est amené à revoir les algorithmes permettant l'estimation et la prédiction du mouvement de l'œil. La thèse CIFRE de Ke Liang qui vient de démarrer s'inscrit dans cette thématique.

4.3 Projets en cours

Concernant les applications en cours, je coordonne les projets Ecriture et Habiletés motrices, et je suis le partenaire responsable de la partie analyse vidéo et de l'intégration des signaux thermiques, physiologiques et comportementaux, du projet ANR Pretherm dont le but est d'obtenir un indicateur fiable du stress et de la douleur chez le bébé prématuré.

Ecriture : En collaboration avec des psychologues de l'Université de Caen de l'équipe PALM, nous travaillons sur un système d'aide à l'apprentissage de l'écriture. Il s'agit de concevoir et d'implémenter une approche de classification et de reconnaissance de lettres. Cette analyse passe par l'extraction de primitives de base d'une lettre. Le but est de mettre en évidence certaines formes de gestes par rapport au modèle du maître. A terme, le projet est destiné à aider les enfants pour l'apprentissage de l'écriture, en analysant et en corrigeant en temps réel le tracé d'une lettre à l'aide de tablettes PC. L'algorithme développé affine la prédiction au cours du traçage.

Habiletés motrices : Projet CPER "STIC et Sécurité" avec pour partenaires : Modesco et l'institut d'Education Motrice F-X Falala qui est un institut financé par l'APF (association des paralysés de France). Dans le cadre de ce projet, nous avons investi dans un cyber glove qui nous permet l'acquisition des mouvements d'une main et leur restitution en environnement 3D. Ce gant nous ouvre de nombreuses perspectives en analyse de

données, en interprétation temps réel de gestes et en analyse de la motricité. Dans le cadre de ce projet, nous avons recruté un ingénieur qui travaille sur les primitives gestuelles et sur leur reconstruction en utilisant les moments spatiotemporels.

Pré-Therm : Projet ANR qui regroupe quatre partenaires : le laboratoire PALM (Psychologie des Actions Langagières et Motrices) de l'Université de Caen, le GREYC, le CHU de Caen et Grey-Soft. Mon implication dans le projet Mesures Thermiques de la Douleur et du Stress chez le Prématuré (Pré-Therm) s'articule autour de l'analyse vidéo et de l'intégration des signaux thermiques, physiologiques et comportementaux dans le but d'obtenir un indicateur fiable du stress et de la douleur chez le bébé prématuré.

Bibliographie

[ABP03] Jürgen Assfalg, Alberto Del Bimbo, and Pietro Pala. Retrieval of 3d objects using curvature maps and weighted walkthroughs. pages 348–353, 2003.

[ACS09] Pablo Arias, Vicent Caselles, and Guillermo Sapiro. A Variational Framework for Non-local Image Inpainting. In Daniel Cremers, Yuri Boykov, Andrew Blake, and Frank R. Schmidt, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 5681, chapter 26, pages 345–358. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[AK06] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing : Partial Differential Equations and the Calculus of Variations (second edition)*, volume 147 of *Applied Mathematical Sciences*. Springer-Verlag, 2006.

[AP02] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. *Lecture notes in computer science*, pages 15–26, 2002.

[AR11] J.K. Aggarwal and M.S. Ryoo. Human activity analysis : A review. *ACM Comput. Surv.*, 43 :16 :1–16 :43, April 2011.

[Azr07] Arik Azran. The rendezvous algorithm : Multiclass semi-supervised learning with markov random walks. 2007.

[BBDB+09] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. November 2009. (Poster).

[BBS01] M. Bertalmio, A. L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. pages 355–362, 2001.

[BCM08] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Nonlocal image and movie denoising. *International Journal of Computer Vision*, 76(2) :123–139, 2008.

[BEM09] Sébastien Bougleux, Abderrahim Elmoataz, and Mahmoud Melkemi. Local and nonlocal discrete regularization on weighted graphs for image and mesh processing. *Int. J. Comput. Vision*, 84 :220–236, August 2009.

[BEV+07] Xavier Bresson, Selim Esedoglu, Pierre Vandergheynst, Jean-Philippe Thiran, and Stanley Osher. Fast global minimization of the active contour/snake model. *J. Math. Imaging Vis.*, 28 :151–167, June 2007.

[BFL06] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient n-d image segmentation. *Int. J. Comput. Vision*, 70 :109–131, November 2006.

[BH09] Thomas Bühler and Matthias Hein. Spectral clustering based on the graph p-laplacian. pages 81–88, 2009.

[BHS09] Martin Burger, Lin He, and Carola-Bibiane Schönlieb. Cahn-hilliard inpainting and a generalization for grayvalue images. *SIAM J. Img. Sci.*, 2 :1129–1167, November 2009.

[Bie87] Irving Biederman. Recognition-by-components : A theory of human image understanding. *Psychological Review*, 94 :115–147, 1987.

[BKB07] Jerome Boulanger, Charles Kervrann, and Patrick Bouthemy. Space-time adaptation for patch-based image sequence restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29 :1096–1102, June 2007.

[BKT06] Matthieu Bray, Pushmeet Kohli, and Philip H. S. Torr. Posecut : Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. pages 642–655, 2006.

[BL93] R. Bergevin and M. D. Levine. Generic object recognition : Building and matching coarse descriptions from line drawings. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15 :19–36, January 1993.

[BMW+10] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Une approche neuronale pour la classification d,Äôactions de sport par la prise en compte du contenu visuel et du mouvement dominant. October 2010.

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15 :1373–1396, 2003.

[BN08] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *J. Comput. Syst. Sci.*, 74(8) :1289–1308, 2008.

[BP94] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. *Advances in Neural Information Processing Systems*, pages 753–753, 1994.

[BSCB00a] Marcelo Bertalmío, Guillermo Sapiro, Vicent Caselles, and Coloma Ballester. Image inpainting. pages 417–424, 2000.

[BSCB00b] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. pages 417–424, 2000.

[Bur10] Christopher J. C. Burges. Geometric methods for feature extraction and dimensional reduction - a guided tour. In *Data Mining and Knowledge Discovery Handbook*, pages 53–82. 2010.

[Can99] N. Canterakis. 3d zernike moments and zernike affine invariants for 3d image analysis and recognition. pages 85–93, 1999.

[CCA09] Guillermo Cámara-Chávez and Arnaldo de Albuquerque Araújo. Harris-sift descriptor for video event detection based on a machine learning approach. In *Proceedings of the 2009 11th IEEE International Symposium on*

Multimedia, ISM '09, pages 153–158, Washington, DC, USA, 2009. IEEE Computer Society.

[CG99] J. Cai and A. Ardeshir Goshtasby. Detecting human faces in color images. *Image Vision Comput.*, 18(1) :63–75, 1999.

[Cha05] Antonin Chambolle. Total variation minimization and a class of binary mrf models. pages 136–152, 2005.

[Chu97] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[CKKS02] Tony F. Chan, Sung Ha Kang, Kang, and Jianhong Shen. Euler's elastica and curvature based inpaintings. *SIAM J. Appl. Math*, 63 :564–592, 2002.

[CKS97] Caselles, R Kimmel, and G Sapiro. Geodesic active contours. *International Journal Of Computer Vision*, 22(1) :61–79, FEB-MAR 1997.

[CPT03] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Object removal by exemplar-based inpainting. pages 721–728, 2003.

[CS02] Tony F. Chan and Jianhong Shen. Mathematical models for local nontexture inpaintings. *SIAM J. Appl. Math*, 62 :1019–1043, 2002.

[CS05a] Tony F. Chan and Jianhong Shen. Image processing and analysis - variational, pde, wavelet, and stochastic methods. pages I–XXI, 1–400, 2005.

[CS05b] Tony F. Chan and Jianhong Shen. Variational image inpainting. *Comm. Pure Applied Math*, 58 :579–619, 2005.

[CSS+11] L. Costantini, L. Seidenari, G. Serra, L. Capodiferro, and A. Del Bimbo. Space-time zernike moments and pyramid kernel descriptors for action classification. pages 199–208, 2011.

[CTL+10] Liangliang Cao, YingLi Tian, Zicheng Liu, Benjamin Yao, Zhengyou Zhang, and Thomas S. Huang. Action detection using multiple spatial-temporal interest point features. pages 340–345, 2010.

[CV01] T. F. Chan and L. A. Vese. Active contours without edges. *Image Processing, IEEE Transactions on*, 10(2) :266–277, February 2001.

[CW97] Cheng-Chia Chang and Ling-Ling Wang. A fast multilevel thresholding method based on lowpass and highpass filtering. *Pattern Recognition Letters*, 18(14) :1469–1478, 1997.

[DB08] Julia A. Dobrosotskaya and Andrea L. Bertozzi. A wavelet-laplace variational technique for image deconvolution and inpainting. *IEEE Transactions on Image Processing*, 17(5) :657–663, 2008.

[DCB] BT David, R. Chalon, and M. Beldame. Oeil et IHM : suivi du regard et interaction" à líoeil.

[DCC+08] Jérôme Darbon, Alexandre Cunha, Tony F. Chan, Stanley Osher, and Grant J. Jensen. Fast nonlocal filtering applied to electron cryomicroscopy. pages 1331–1334, 2008.

[DD10] Julie Delon and Agnès Desolneux. Stabilization of flicker-like effects in image sequences through local contrast correction. *SIAM J. Imaging Sciences*, 3(4) :703–734, 2010.

[DDT09] Charles-Alban Deledalle, Loïc Denis, and Florence Tupin. Iterative weighted maximum likelihood denoising with probabilistic patch-based weights. *Trans. Img. Proc.*, 18 :2661–2672, December 2009.

[DFE07] Kostadin Dabov, Alessandro Foi, and Karen Egiazarian. Video denoisong by sparse 3d transform-domain collaborative filtering. *Signal Processing*, 1(Eusipco) :145–149, 2007.

[DNL09] 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.

[DRCB05] Piotr Dollar, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. pages 65–72, 2005.

[Duc07] A.T. Duchowski. *Eye tracking methodology : Theory and practice*. Springer-Verlag New York Inc, 2007.

[EL99] Alexei Efros and Thomas Leung. Texture synthesis by non-parametric sampling. pages 1033–1038, 1999.

[ELB08] Abderrahim Elmoataz, Olivier Lezoray, and Sébastien Bougleux. Nonlocal discrete regularization on weighted graphs : A framework for image and manifold processing. *IEEE Transactions on Image Processing*, 17(7) :1047–1060, 2008.

[ELBT08] A. Elmotaz, O. Lezoray, S. Bougleux, and V. Ta. Unifying local and nonlocal processing with partial difference operators on weighted graphs. pages 11 – 26, 2008.

[Fid85] M. A. Fiddy. The Radon Transform and Some of Its Applications (book revision). *Journal of Modern Optics*, 32(1) :3–4, 1985.

[FZ05] D. Freedman and Tao Zhang. Interactive graph cut based segmentation with shape priors. 1 :755–762, 2005.

[GAM+06] Liwei Guo, Oscar C. Au, Mengyao Ma, Zhiqin Liang, and Carman K. M. Yuk. A multihypothesis motion-compensated temporal filter for video denoising. pages 1417–1420, 2006.

[GBS+05] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. pages 1395–1402, 2005.

[GEL11] Mahmoud Ghoniem, Abderrahim Elmoataz, and Olivier Lezoray. Discrete infinity harmonic functions : Towards a unified interpolation framework on graphs. pages 1361–1364, 2011.

[GHS11] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. A time series kernel for action recognition. In *British Machine Vision Conference*, Dundee, United Kingdom, August 2011.

[GO07a] Guy Gilboa and Stanley Osher. Nonlocal linear image regularization and supervised segmentation. *Multiscale Modeling & Simulation*, 6(2) :595–630, 2007.

[GO07b] Guy Gilboa and Stanley Osher. Nonlocal linear image regularization and supervised segmentation. 6(2) :595–630, 2007.

[GO08] Guy Gilboa and Stanley Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3) :1005–1028, 2008.

[Gru08] M. Grunwald. *Human haptic perception : basics and applications*. 2008.

[GSS02] G. Gomez, M. Sanchez, and Luis Enrique Sucar. On selecting an appropriate colour space for skin detection. pages 69–78, 2002.

[GZ06] I Gutman and B Zhou. Laplacian energy of a graph. *Linear Algebra and its Applications*, 414(1) :29–37, 2006.

[HAMJ02] Rein-Lien Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5) :696–706, 2002.

[HGBRM10] Nguyen Vu Hoàng, Valérie Gouet-Brunet, Marta Rukoz, and Maude Manouvrier. Embedding spatial information into image content description for scene retrieval. *Pattern Recognition*, 43(9) :3013–3024, 2010.

[HM07] Matthias Hein and Markus Maier. Manifold denoising as preprocessing for finding natural representations of data. pages 1646–1649, 2007.

[HNL06] C. Hennessey, B. Noureddin, and P. Lawrence. A single camera eye-gaze tracking system with free head motion. pages 87–94, 2006.

[HS81] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *ARTIFICAL INTELLIGENCE*, 17 :185–203, 1981.

[JBBA01a] S. Jehan-Besson, M. Barlaud, and G. Aubert. Contours actifs basés régions pour la segmentation des objets en mouvement dans les séquences à caméra fixe ou mobile. 2001.

[JBBA01b] Stéphanie Jehan-Besson, Michel Barlaud, and Gilles Aubert. Video objects segmentation using eulerian region-based active contours. pages 353–361, 2001.

[JR02a] Michael J. Jones and James M. Rehg. Compaq skin database. 2002.

[JR02b] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46 :81–96, January 2002.

[KB08] Charles Kervrann and Jérôme Boulanger. Local adaptivity to variable smoothness for exemplar-based image regularization and representation. *Int. J. Comput. Vision*, 79 :45–69, August 2008.

[kKfWC07] Tae kyun Kim, Shu fai Wong, and Roberto Cipolla. R. : Tensor canonical correlation analysis for action classification. 2007.

[KMLS10] Alexander Kläser, Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Will person detection help bag-of-features action recognition ? Rapport de recherche RR-7373, INRIA, France, September 2010.

[KOJ05] Stefan Kindermann, Stanley Osher, and Peter W Jones. Deblurring and denoising of images by nonlocal functionals. *Multiscale Modeling Simulation*, 4(4) :1091, 2005.

[Kon94] I. Kononenko. Estimating Attributes : Analysis and Extensions of RELIEF. *Proc. of the European Conference on Machine Learning, ECML*, pages 171–182, 1994.

[KPVG10] Jan Knopp, Mukta Prasad, and Luc Van Gool. Orientation invariant 3D object classification using hough transform based methods. pages 15–20, 2010.

[KT07] Pushmeet Kohli and Philip H. S. Torr. Dynamic graph cuts for efficient inference in markov random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29 :2079–2088, December 2007.

[KTC03] Elena Karpova, Dzmitry Tsishkou, and Liming Chen. The ECL Skin-color Images from Video (SCIV) Database, 2003. in Proceeding of IAPR International Conference on Image and Signal Processing (ICISP'2003).

[KVV04] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings : Good, bad and spectral. *J. ACM*, 51 :497–515, May 2004.

[KWT88] Michael Kass, Andrew P. Witkin, and Demetri Terzopoulos. Snakes : Active contour models. *International Journal of Computer Vision*, 1(4) :321–331, 1988.

[KZ04] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts ? *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2) :147–159, 2004.

[KZM11] V. Kellokumpu, G Zhao, and Pietikainen M. Recognition of human actions using texture descriptors. *Machine Vision and Applications, in press (available online)*, 2011.

[KZP11] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Recognition of human actions using texture descriptors. *Mach. Vis. Appl.*, 22(5) :767–780, 2011.

[LBS09] J. Lellmann, F. Becker, and C. Schnörr. Convex optimization for multi-class image labeling with a novel family of total variation based regularizers. pages 646 – 653, 2009.

[LHDN10] Dijun Luo, Heng Huang, Chris Ding, and Feiping Nie. On the eigenvectors of p-laplacian. *Mach. Learn.*, 81 :37–51, October 2010.

[LK81] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. pages 674–679, 1981.

[LK87] Susan J Lederman and Roberta L Klatzky. Hand movements : A window into haptic object recognition. *Cognitive Psychology*, 19(3) :342 – 368, 1987.

[LL06] Stephane Lafon and Ann B. Lee. Diffusion maps and coarse-graining : A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 :1393–1403, 2006.

[LM02] R. Lienhart and J. Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. 2(1) :900–903, 2002.

[Luo06] *Non-Local Image Interpolation*, 2006.

[LV07] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Information Science and Statistics Series. Springer, 2007.

[LWP05] D. Li, D. Winfield, and D.J. Parkhurst. Starburst : A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. pages 1–8, 2005.

[MB09] A Massonneau and N Biard. Etat de l'art des différents systèmes de pointages à l'oeil. *Plate-Forme Nouvelles Technologies*, 2009.

[MKAF99] CH Morimoto, D. Koons, A. Amir, and M. Flickner. Frame-rate pupil detector and gaze tracker. 99, 1999.

[MM96] B. S. Manjunath and W. Y. Ma. Texture Features for Browsing and Retrieval of Image Data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8) :837–842, August 1996.

[MM98] Simon Masnou and Jean-Michel Morel. Level lines based disocclusion. pages 259–263, 1998.

[MP02] Etienne Mémin and Patrick Pérez. Hierarchical estimation and segmentation of dense motion fields. *Int. J. Comput. Vision*, 46 :129–155, February 2002.

[MS89] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42 :577 – 685, 1989.

[MS01] Marina Maila and Jianbo Shi. A random walks view of spectral segmentation. 2001.

[MSP03] Birgitta Martinkauppi, Maricor Soriano, and Matti Pietikäinen. Detection of skin color under changing illumination : A comparative study. pages 652–, 2003.

[MYC+10] Susan McKeever, Juan Ye, Lorcan Coyle, Chris J. Bleakley, and Simon Dobson. Activity recognition using temporal evidence theory. *JAISE*, 2(3) :253–269, 2010.

[NEC06] Mila Nikolova, Selim Esedoglu, and Tony F. Chan. Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models. *SIAM Journal on Applied Mathematics*, 66(5) :1632–1648, 2006.

[New06] The Structure and Dynamics of Networks. 2006.

[NHW+08] Huazhong Ning, Tony X. Han, Dirk B. Walther, Ming Liu, and Thomas S. Huang. Hierarchical space-time model enabling efficient search for human actions, 2008.

[Nik07] Vladimir Nikiforov. The energy of graphs and matrices. *Journal of Mathematical Analysis and Applications*, 326(2) :1472 – 1475, 2007.

[NK03] Marcin Novotni and Reinhard Klein. 3d zernike descriptors for content based shape retrieval. pages 216–225, 2003.

[NWFf06] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-fei. Unsupervised learning of human action categories using spatial-temporal words. 2006.

[OK06] Ryutarou Ohbuchi and Jun Kobayashi. Unsupervised learning from a corpus for shape-based 3d model retrieval. pages 163–172, 2006.

[OM04] T. Ohno and N. Mukawa. A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. pages 115–122, 2004.

[OMK03] T. Ohno, N. Mukawa, and S. Kawato. Just blink your eyes : a head-free gaze tracking system. pages 950–957, 2003.

[OMY02] T. Ohno, N. Mukawa, and A. Yoshikawa. FreeGaze : a gaze tracking system for everyday gaze interaction. pages 125–132, 2002.

[PAF+09] Roberto Perdisci, Davide Ariu, Prahlad Fogla, Giorgio Giacinto, and Wenke Lee. Mcpad : A multiple classifier system for accurate payload-based anomaly detection. *Comput. Netw.*, 53 :864–881, April 2009.

[PBC08] Gabriel Peyré, Sébastien Bougleux, and Laurent Cohen. Non-local regularization of inverse problems. pages 57–68, 2008.

[PCBC10] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. 3 :1122–1145, 2010.

[PCG+03] A. Perez, ML Cordoba, A. Garcia, R. Mendez, ML Munoz, JL Pedraza, and F. Sanchez. A precise eye-gaze detection and tracking system. 2003.

[PD00] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :266–280, March 2000.

[PD02] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 46 :223–247, 2002.

[PD05] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for motion estimation and tracking. *Computer Vision and Image Understanding*, 97(3) :259–282, March 2005.

[PE07] Matan Protter and Michael Elad. Sparse and redundant representations and motion-estimation-free algorithm for video denoising. 2007.

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572, 1901.

[PGM10] Dinesh J. Peter, V. K. Govindan, and Abraham T. Mathew. Nonlocal-means image denoising technique using robust m-estimator. *J. Comput. Sci. Technol.*, 25 :623–631, May 2010.

[PMD+10] Luis Pizarro, Pavel Mrázek, Stephan Didas, Sven Grewenig, and Joachim Weickert. Generalised nonlocal image smoothing. *International Journal of Computer Vision*, 90(1) :62–87, 2010.

[PR99] Eric Paquet and Marc Rioux. The mpeg-7 standard and the content-based management of three-dimensional data : A case study. pages 375–380, 1999.

[PSG+08] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. A convex formulation of continuous multi-label problems. October 2008.

[Qui86] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1 :81–106, March 1986.

[Qui93] Ross J. Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[Ras06] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[RD04] F. Ranchin and F. Dibos. Moving objects segmentation using optical flow estimation. *Mathematics, Image and Analysis*, 2006, 2004.

[RDLTH01] G Robles-De-La-Torre and V Hayward. Force can overcome object geometry in the perception of shape through active touch. *Nature*, 412(6845) :445–448, 2001.

[RLB09] Jérome Revaud, Guillaume Lavoué, and Atilla Baskurt. Improving zernike moments comparison for optimal similarity and rotation angle retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4) :627–636, 2009.

[ROF92a] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D Nonlinear Phenomena*, 60(1-4) :259–268, November 1992.

[ROF92b] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60 :259–268, November 1992.

[RRS00] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438. ACM New York, NY, USA, 2000.

[SAK11] Hichem Sahbi, Jean-Yves Audibert, and Renaud Keriven. Context-Dependent Kernels for Object Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4) :699 – 708, April 2011.

[SBdHK11] Dan Schonfeld, Jan Biemond, Gerard de Haan, and André Kaup. Introduction to the issue on recent advances in video processing for consumer displays. *J. Sel. Topics Signal Processing*, 5(2) :213–216, 2011.

[Sha67] B. Shackel. Eye movement recording by electro-oculography. *A manual of psychophysiological methods*, pages 299–334, 1967.

[She06] J. J. S. Shen. A stochastic-variational model for soft mumford-shah segmentation. *International Journal of Biomedical Imaging*, 2006, 2006.

[SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :888–905, 2000.

[SMC08] Arthur D. Szlam, Mauro Maggioni, and Ronald R. Coifman. Regularization on graphs with function-adapted diffusion processes. *J. Mach. Learn. Res.*, 9 :1711–1739, June 2008.

[SN01] Jamie D. Shutler and Mark S. Nixon. Zernike velocity moments for description and recognition of moving shapes. pages 705–714, 2001.

[Spe04] Charles Spearman. "General Intelligence," Objectively Determined and Measured. *American Journal of Psychology*, 15 :201–293, 1904.

[SV99] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9 :293–300, 1999. 10.1023/A :1018628609742.

[SW00] Christoph Schnörr and Joachim Weickert. Variational image motion computation : Theoretical framework, problems and perspectives. pages 476–488, 2000.

[SWH+06] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee. Spectral methods for dimensionality reduction. *Semisupervised Learning. MIT Press : Cambridge, MA*, 2006.

[SYW97] R. Stiefelhagen, J. Yang, and A. Waibel. Tracking eyes and monitoring eye gaze. pages 98–100, 1997.

[TD05] David Tschumperlé and Rachid Deriche. Vector-valued image regularization with pdes : A common framework for different applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(4) :506–517, 2005.

[THES06] Alireza Tavakoli Targhi, Eric Hayman, Jan-Olof Eklundh, and Mehrdad Shahshahani. The eigen-transform and applications. pages 70–79, 2006.

[TJW01] Andy Tsai, Anthony J. Yezzi Jr., and Alan S. Willsky. Curve evolution implementation of the mumford-shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Transactions on Image Processing*, pages 1169–1186, 2001.

[TM98] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. pages 839–, 1998.

[TSL00] Joshua B. Tenenbaum, Vin Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500) :2319–2323, December 2000.

[TV08] Johan W. Tangelder and Remco C. Veltkamp. A survey of content based 3d shape retrieval methods. *Multimedia Tools Appl.*, 39 :441–471, September 2008.

[TWL+10a] Anh-Phuong Ta, Christian Wolf, Guillaume Lavoué, Atilla Baskurt, and Jean-Michel Jolion. Pairwise features for human action recognition. pages 3224–3227, 2010.

[TWL+10b] Anh-Phuong Ta, Christian Wolf, Guillaume Lavoué, Atilla Baskurt, and Jean-Michel Jolion. Pairwise features for human action recognition. pages 3224–3227, 2010.

[UPL10] Muhammad Muneeb Ullah, Sobhan Naderi Parizi, and Ivan Laptev. Improving bag-of-features action recognition with non-local cues. In *Proceedings of the British Machine Vision Conference*, pages 95.1–95.11. BMVA Press, 2010. doi :10.5244/C.24.95.

[VBC10] R. Vezzani, D. Baltieri, and R. Cucchiara. Hmm based action recognition with projection histogram features. pages 286–293, 2010.

[VCTC02] Luminita A. Vese, Tony F. Chan, Tony, and F. Chan. A multiphase level set framework for image segmentation using the mumford and shah model. *International Journal of Computer Vision*, 50 :271–293, 2002.

[vdMPvdH07] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality Reduction : A Comparative Review. 2007.

[vdSGS10] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9) :1582–1596, 2010.

[VJ01] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. 1, 2001.

[WhLZH06] Zhili Wu, Chun hung Li, Ji Zhu, and Jian Huang. A semi-supervised svm for manifold learning. pages 490–493, 2006.

[Win06] Gerhard Winkler. Image analysis, random fields and markov chain monte carlo methods : A mathematical introduction (stochastic modelling and applied probability). 2006.

[WKC07] Shu-Fai Wong, Tae-Kyun Kim, and Roberto Cipolla. Learning motion categories using both semantic and structural information. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[WO08] Alexander Wong and Jeff Orchard. A nonlocal-means approach to exemplar-based inpainting. pages 2600–2603, 2008.

[WWL07] Fei Wang, Xin Wang, and Tao Li. Efficient label propagation for interactive image segmentation. pages 136–141, 2007.

[WXDL11] Xinxiao Wu, Dong Xu, Lixin Duan, and Jiebo Luo. Action recognition using context and appearance distribution features. pages 489–496, 2011.

[XHZX12] Dong Xu, Yi Huang, Zinan Zeng, and Xinxing Xu. Human gait recognition using patch distribution feature and locality-constrained group sparse representation. *IEEE Transactions on Image Processing*, 21(1) :316–326, 2012.

[XMH09] Sun. Xinghua, Chen. Mingyu, and A. Hauptmann. Action recognition via local descriptors and holistic features. *Computer Vision and Pattern Recognition Workshop*, 0 :58–65, 2009.

[YC05] D.H. Yoo and M.J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 98(1) :25–51, 2005.

[YKA02] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja. Detecting faces in images : A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(1) :34–58, 2002.

[YS75] L.R. Young and D. Sheena. Survey of eye movement recording methods. *Behavior research methods and instrumentation*, 7(5) :397–429, 1975.

[zar99] *Comparison of five color models in skin pixel classification*, 1999.

[ZJ05] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. 1, 2005.

[ZL01] Dengsheng Zhang and Guojun Lu. Segmentation of moving objects in image sequence : A review. *Circuits, Systems, and Signal Processing*, 20 :143–183, 2001. 10.1007/BF01201137.

[ZmI01] L. Zelnik-manor and M. Irani. Event-based analysis of video. pages 123–130, 2001.

[ZP03] Titus B. Zaharia and Françoise J. Prêteux. Descripteurs de forme pour l'indexation de maillages 3d. *Technique et Science Informatiques*, 22(9) :1077–1105, 2003.

[ZPP06] V. Zlokolica, A. Pizurica, and W. Philips. Wavelet-domain video denoising based on reliability measures. 16(8) :993–1007, August 2006.

[ZR00] D. Zighed and R. Rakotomalala. *Graphes d'Induction : Apprentissage et Data Mining*. Hermes, 2000.

[ZS04] Dengyong Zhou and Bernhard Scholkopf. Learning from labeled and unlabeled data using random walks. pages 237–244, 2004.

[ZS05] Dengyong Zhou and Bernhard Schölkopf. Regularization on discrete spaces. pages 361–368, 2005.

[ZWBL11] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic Imaging*, 20(2) :023016–16, April 2011.

[ZWM97] Song Chun Zhu, Ying Nian Wu, and David Mumford. Minimax entropy principle and its application to texture modeling. *Neural Comput.*, 9 :1627–1660, November 1997.

[ZY96] Song Chun Zhu and Alan Yuille. Region competition : Unifying snakes, region growing, and bayes/mdl for multi-band image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18 :884–900, 1996.

Contributions à la Caractérisation & Structuration d'images et de séquences vidéo

Résumé :

Le mémoire présente une synthèse des travaux en analyse de séquences d'images et de vidéo.

En première partie, je présente mes travaux en régularisation spatio-temporelle basée graphes pour résoudre des problématiques ouvertes en traitement des images : la restauration de séquences d'images bruitées (débruitage) et données manquantes (inpainting), ainsi que la classification de données dans des séquences vidéos. L'approche de régularisation utilise une famille d'opérateurs basée sur le laplacien et pour la classification nous exploitons l'espace intrinsèque du graphe obtenu par une réduction non linéaire de dimension de l'espace d'analyse grâce aux marches aléatoires.

La seconde partie est consacrée à l'analyse de séquences vidéo. Dans un premier temps, je présente notre modèle de peau hybride basé sur les graphes d'induction ainsi que mes travaux liés à la segmentation convexe en volume d'un objet en mouvement à partir des contours actifs et du flot optique. Un autre volet est consacré à des travaux, issus de projets pluridisciplinaires, liés à l'analyse de gestes et du regard. Les bases de vidéos d'actions spécifiques sont analysées en utilisant la régularisation discrète sur graphe et les moments spatio-temporels comme descripteurs de vidéos. Enfin, je présente mes travaux concernant l'analyse du regard en temps réel par processus gaussiens.