



Master 2 recherche SIC
Rapport de Stage

# Tenseurs pour le calcul du flot optique et la description des actions dans vidéos

Laboratoire

Équipes Traitement de l'Information et Systèmes (ETIS)

Présenté par : Virgínia FERNANDES MOTA

Encadré par : Philippe-Henri GOSSELIN
Collaboration : Frédéric PRECIOSO
Marcelo BERNARDES VIEIRA
Sylvie PHILIPP-FOLIGUET



Septembre, 2010

# Remerciements

Tout d'abord, je souhaite remercier Phillip-Henri GOSSELIN, Frédéric PRECIOSO, Marcelo BERNARDES VIEIRA et Sylvie PHILIPP-FOLIGUET pour leur encadrement et pour leur patience. Un grand merci également à toute l'équipe du Master SIC. Merci à ENSEA, à UCP et à UFJF pour l'opportunité et le financement. Je remercie aussi mes amis, tous les membres de ma famille et mon fiancé Tiago pour leur soutien.

## Résumé

Dans ce travail, nous proposons un descripteur global pour les actions humaines de la base de vidéos KTH basé sur le tenseur d'orientation. Ce tenseur d'orientation est composé à partir des coefficients des polynômes de Legendre calculés pour chaque *frame* d'un vidéo. Nous avons tester le descripteur créé avec un classifieur SVM. Nous voyons que la précision de notre approche est encore loin des précisions présentent dans la littérature, par contre, le descripteur se montre prometteur.



## Abstract

In this work, we propose a global descriptor for human actions of the video database KTH based on the orientation tensor. The orientation tensor is composed by the coefficients of Legendre polynomials calculated for each frame of a video. We test our descriptor with a SVM classifier. We see that the accuracy of our approach is still far from the one present in literature, but the descriptor show to be promising.

# Table des matières

# Chapitre 1

# Introduction

La recherche en Vision par Ordinateur s'applique à définir des algorithmes permettant la perception et la compréhension du monde physique à partir d'informations visuelles (images et séquences d'images). Un système de vision complet s'articule autour de trois étapes : l'extraction d'attributs de bas-niveau (couleur, texture, forme, orientation, mouvement,...), l'analyse des attributs, fournissant des informations de plus haut-niveau sur la scène (reconnaissance, segmentation, catégorisation), et l'interprétation de la scène. Le travail présenté dans ce rapport de stage a pour sujet la reconnaissance des actions dans les vidéos.

Le mouvement constitue une source d'information visuelle très importante. Il offre des informations sur la structure tridimensionnelle de la scène, la trajectoire des objets, l'activité qui se déroule dans la scène. L'estimation de toutes ces informations nécessite une mesure précise et pertinente du mouvement dans les séquences d'images.

Une des techniques d'estimation locale du mouvement est le calcul du flot optique. Le flot optique est le champ de vitesse (ou déplacement) décrivant le mouvement apparent des motifs d'intensité de l'image sous l'hypothèse d'illumination constante. Une des approches actuelles pour l'extraction du flot optique est d'utiliser le tenseur de structure qui décrit une information locale moyennée des orientations et préserve la structure du mouvement [1],[2].

Ces méthodes basées sur les tenseurs estiment généralement le mouvement à partir d'approches différentielles et de filtrage, puisque le mouvement peut être caractérisé comme la variation de l'intensité des pixels entre les images adjacentes. Le tenseur de structure permet d'évaluer la covariance entre les frames successives en décrivant la structure spatio-temporelle d'un voisinage donné. De plus, il fournit l'information par rapport à la vitesse locale, c'est-à-dire la valeur du vrai flot (le flot optique normal).

Dès que nous avons l'information du mouvement de la vidéo, nous pouvons passer à la deuxième étape du système de vision, l'étape d'analyse. Une méthode très intéressante pour l'analyse du mouvement c'est la modélisation par des combinaison linéaires de polynômes orthogonaux. Cette méthode permet d'obtenir

une expression polynomiale du mouvement [3],[4].

Nous proposons donc un nouveau descripteur vidéo basé sur le tenseur d'orientation composé par les coefficients de polynômes orthogonaux calculés à partir de la séquence d'images.

Nous avons travaillé sur la base de données vidéos KTH [5] où nous avons six types d'actions humaines (*walking*, *jogging*, *running*, *boxing*, *hand waving* et *hand clapping*) faites plusieurs fois par 25 sujets différents dans quatre scénarios : dehors (s1), dehors avec variation d'échelle (s2), dehors avec différents vêtements (s3) et à l'intérieur (s4) (FIGURE 1.1). Toutes les séquences ont été prises avec un fond homogène et une caméra statique de 25fps. Les séquences ont une résolution de 160x120 pixels et durent environ quatre secondes.

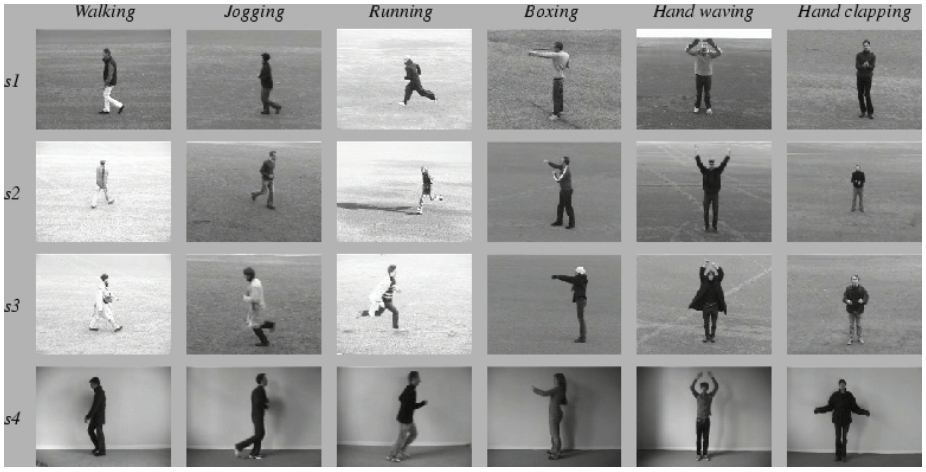



FIGURE 1.1 – Types d'actions

# Chapitre 2

# Étude Bibliographique

Dans ce chapitre, nous dressons un état de l'art des différentes méthodes permettant d'extraire le mouvement contenu dans des séquences d'images et de les analyser.

L'approche de Riemenschneider [6] considère la séquence d'images comme un volume spatio-temporel et détecte les *Maximally Stable Volumes* (MSVs) dans les champs de flot optique. La méthode pour identifier les séquences d'images se compose de deux étapes : détection et description des points 3D d'intérêt (*feature vectors*) et *standard bag-of-words model* (décrit les séquences d'image en termes de signature du volume flot optique).

La détection et description des points 3D d'intérêt se compose principalement de trois étapes : l'estimation du flot optique (par la méthode TV-L1 [7]), l'application du détecteur *Maximally Stable Volume* (MSV) pour identifier les volumes stables du flot optique et la description des points d'intérêt situés sur les surfaces du volume avec un descripteur 3D analysant la forme locale des volumes. Les *visual words* sont identifiés en groupant l'ensemble de tous les descripteurs pour caractériser leurs propriétés partagées. À partir du nombre d'occurrences des *visual words* dans la séquence d'images, une signature distincte est construite.

La méthode de cet article est très riche et peut être utilisée pour tout types de mouvements, même mouvements complexes du corps.

Dans [8] la méthode est basée dans l'OFT–HMMs framework. Dans ce framework, les unités sémantiques sont prédéfinies selon la connaissance spécifique du domaine de l'application pratique, et les échantillons d'apprentissage sont choisis pour la modélisation de HMMs (Hidden Markov Model). Il y a deux étapes principales : extraction du flot optique (OFT) et l'application de HMM.

Pour faire l'extraction du flot optique, la méthode de Lucas et Kanade est particulièrement attrayante pour ses attributs notables tels que l'exécution de grande précision et l'efficacité, mais, comme c'est une approche basée-gradient, la méthode de Lucas et Kanade se trompe parfois face au problème de grand mouvement (*large motion problem*). Pour résoudre ce problème, ils ont utilisé la méthode hierarchique Lucas–Kanade (HLK). On a un ensemble d'images

dérivées des images vidéo originales à plusieurs résolutions comme une structure hiérarchisée de pyramide. Le flot optique approximé est d'abord calculé entre les images avec la plus basse résolution, et les plus précises sont obtenues par le calcul entre les images avec une résolution plus élevée.

Afin de préserver l'information distincte des tenseurs d'apprentissage, le flot optique est analysé par la méthode *general tensor discriminant analysis* (GTDA) et *linear discriminant analysis* (LDA), et produit les vecteurs finaux. Comme le HMM est un outil d'analyse séquentielle efficace, les HMMs sont adoptés pour modéliser les unités sémantiques, tels que les plans et les actions.

Une méthode similaire est présentée dans [9], sauf que pour analyser les tenseurs il n'utilise pas le GTDA et LDA, mais il utilise LDA et PCA (*principal component analysis*).

Les méthodes proposées dans [8] et [9] sont aussi très riches et peuvent être utilisées pour les mouvements complexes.

Dans [4], la méthode présentée propose de caractériser tout type de mouvement comme une combinaison linéaire de polynômes issus d'une base orthonormée. À partir de la séquence originale, tous les champs de vecteurs sont extraits par une méthode différentielle d'estimation du mouvement apparent. Cette méthode est fondée sur l'hypothèse de la conservation de la luminance des pixels dans une image et sur l'utilisation d'équations aux dérivées partielles (E.D.P.) de lissage directionnel. Celles-ci utilisent un opérateur différentiel, appelé tenseur de structure, permettant de déterminer localement, à partir de données spatio-temporelles, la direction du mouvement apparent. Cette direction est par ailleurs utilisée pour lisser la séquence d'images dans la direction du mouvement. Une évaluation du champ dense et régularisé est ainsi obtenue. Pour chacun de ces champs, les polynômes caractéristiques sont alors calculés par projections sur la base. Finalement, le mouvement est déterminé en étudiant les variations dans le temps des coefficients de ces polynômes.

La méthode de cet article est très simple, mais elle est présentée seulement pour les mouvements du visage.

Dans [3] un étude plus approfondi de la modélisation du mouvement par base de polynômes est proposé, comme trouvée en [4], mais appliquée à l'écoulement de fluides. La méthode proposée permet de modéliser, de façon globale, tout type de mouvement par des combinaisons linéaires de polynômes orthogonaux. Ce modèle est évalué sur trois séquences expérimentales. Les résultats obtenus sont comparés à une décomposition orthogonale aux valeurs propres. Cette méthode permet de modéliser, pour chaque séquence, plus de 99% de l'énergie cinétique de l'écoulement, ce qui permet d'avoir une bonne représentation des structures principales. Cette méthode également a permis d'obtenir une expression polynomiale du mouvement étudié. En étudiant le comportement des coefficients de projection, il est possible d'extraire certaines informations.

Tous ces articles cités ci-dessus utilisent leurs propres bases de vidéos. Les trois articles prochains utilisent la même base de vidéo choisie pour ce travail, la base KTH, introduite par [5].

La méthode proposée dans [5] utilise les caractéristiques espace-temps locales. Après l'extraction de ces caractéristiques, ils utilisent un classifieur SVM

(*support vectors machines* ou machines à vecteurs de support) pour déterminer les actions de chaque shot.

Dans [10], nous avons une représentation compacte pour la reconnaissance d'actions humaines dans les vidéos en utilisant les histogrammes de ligne et les histogrammes de flot optique. Ils ont créé un descripteur basé sur la distribution de lignes qui sont formé par les frontières de figures humaines. Ils utilisent également une représentation compacte du flot optique pour avoir des informations de mouvement. Après l'extraction de ces caractéristiques, ils utilisent un classifieur SVM, de la même façon que [5].

Une nouvelle proposition de descripteur peut être trouvée dans [11]. Ils proposent une méthode en utilisant une combinaison efficace d'un nouveau descripteur de gradient 3D avec un descripteur de flot optique, pour représenter les points d'intérêt spatio-temporelle. Ces points sont utilisés pour représenter des séquences vidéo à l'aide d'un ensemble de mots visuels spatio-temporelles. Les classifieurs SVM ont aussi été utilisés pour faire la classification.

Il est intéressant de remarquer que pour ces trois articles ([5], [10] et [11]), les mouvements *walking*, *jogging* et *running* de la base de vidéo KTH présentent des confusions de classification.

En général, les descripteurs pour la reconnaissance d'actions dans les vidéos qui sont composés pour plus d'un type de caractéristique donnent les meilleurs résultats.

# Chapitre 3

# Fondements

## 3.1 Le flot optique

L'estimation du mouvement consiste à mesurer la projection 2D dans le plan de l'image d'un mouvement réel 3D. Le mouvement 2D est aussi appelé flot optique, on peut alors le définir comme le champ de vitesse décrivant le mouvement apparent des motifs d'intensités de l'image sous l'hypothèse d'illumination constante. Une séquence d'images peut être représentée par sa fonction de luminance I(x, y, t). L'hypothèse de conservation de la luminance signifie que la luminance d'un point physique de la séquence d'image ne varies pas au cours du temps, c'est à dire :

$$I(x, y, t) = I(x + dx, y + dy, t + 1) \tag{3.1}$$

Étant donné $D$ le domaine spatio-temporel correspondant à la séquence $I(x, y, t) : D \rightarrow R$, $\Omega$ le domaine spatial de $(x, y)$ et $v(x, y, t) : D \rightarrow R^2$ le champ de vitesse instantané au temps $t$. Le problème est de trouver $v$ au temps $t$. En prenant compte l'hypothèse ci-dessus, nous avons le problème de Contrainte du Flot Optique (CFO) représentée par :

$$\nabla I \ . \ v + I_t \ = \ \nabla I \ . \ \tilde{v} = 0 \tag{3.2}$$

où $\tilde{v} = (v_1, v_2, 1)^T$

Cette équation ne permet que d'obtenir la composante $v_\perp$, parallèle à $\hat{n} = \frac{\nabla I}{\|\nabla I\|}$, c'est le *problème d'ouverture*. On n'accède au mouvement apparent d'un point que grâce à un calcul effectué dans un voisinage borné de ce point. On ne peut calculer la composante du mouvement que dans la direction du gradient (i.e. perpendiculaire au contour). De plus, l'estimation est impossible dans le cas où $\nabla I = 0$.

Admettons qu'on regarde une surface en mouvement (le rectangle) au travers de petites fenêtres (symbolisées par les cercles), on peut avoir trois cas possibles :

– lorsque le gradient d'intensité est nul, le mouvement n'est pas perceptible.

– lorsque le gradient d'intensité dans la fenêtre est orienté dans une seule direction, le mouvement est perçu comme normal au contour.
– la combinaison d'informations issues de gradients de différentes orientations permet de remonter au mouvement réel de l'objet.

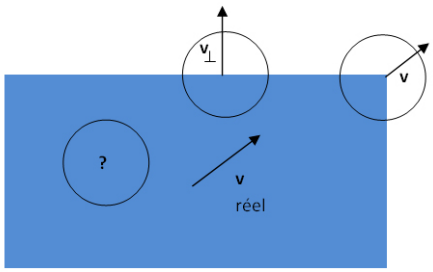



FIGURE 3.1 – Illustration du problème d'ouverture

L'estimation locale du mouvement n'est possible que par l'addition de contraintes issues d'un voisinage spatial ou spatio-temporel, la solution obtenue est alors une moyenne du mouvement sur ce voisinage. Ce voisinage doit être suffisamment large pour contraindre la solution, sans recouvrir de régions contenant des mouvements différents (la solution serait alors une moyenne de tous ces mouvements). Ainsi, en plus de la conservation de la luminance, une hypothèse de continuité spatiale du flot optique est nécessaire pour déterminer le mouvement.

Un exemple classique de manifestation du problème de l'ouverture est l'enseigne de barbier (FIGURE 3.2), pour laquelle on a l'impression que les lignes se déplacent vers le haut, alors qu'elles se déplacent vers la droite.

FIGURE 3.2 – L'enseigne de barbier

## 3.2 Modélisation des champs de déplacement par bases de polynômes

Nous pouvons définir un champ de déplacement (ou flot optique) $F$ de la façon suivante :

$$F : \begin{array}{l} \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2 \\ (x1, x2) \mapsto (V^1(x_1, x_2), V^2(x_1, x_2)) \end{array} \tag{3.3}$$

avec $V^1 : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ et $V^2 : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ deux applications correspondant respectivement aux déplacements horizontaux et verticaux aux points de coordonnées $(x_1, x_2) \in \Omega$.

Nous souhaitons approximer des fonctions réelles à deux variables par des fonctions polynomiales. Nous utilisons donc des polynômes définis dans $\mathbb{R}[x1, x2]$ de la façon suivante :

$$P_{K,L} = \sum_{k=0}^{K} \sum_{l=0}^{L} c_{k,l} (x_1)^k (x_2)^l \tag{3.4}$$

où $K \in \mathbb{N}$ est le degré maximal selon $x_1$, $L \in \mathbb{N}^*$ est le degré maximal selon $x_2$ et $c_{k,l}$ est l'ensemble des coefficients réels du polynôme. Le degré du polynôme est alors $K + L$.

### 3.2.1 Génération d'une base orthogonale bidimensionnelle

Nous pouvons générer une famille de fonctions orthogonales à partir de la formule de récurrence à trois termes suivante :

$$\begin{cases} P_{-1,j} = 0 \\ P_{i,-1} = 0 \\ P_{0,0} = 1 \\ P_{i+1,j} = (a_i x_1 + b_i) P_{i,j} - c_i P_{i-1,j} \\ P_{i,j+1} = (a_j x_1 + b_j) P_{i,j} - c_j P_{i,j-1} \end{cases} \tag{3.5}$$

À partir des valeurs de $a_n$, $b_n$ et $c_n$, nous pouvons générer différentes familles de polynômes orthogonaux connus. Ces polynômes sont alors orthogonaux deux à deux sur le domaine $\Omega$ par rapport au produit scalaire défini par 3.6 relativement à la fonction de poids $\omega(x_1, x_2)$.

$$\langle f_1(x) | f_2(x) \rangle = \int_\Omega f_1 f_2 \omega(x) dx \tag{3.6}$$

Notre base bidimensionnelle $B = P_{i,j}$ est donc composée de polynômes orthonormaux. Nous appelons degré $g$ de cette base le degré le plus élevé des polynômes qui la composent. Une base bidimensionnelle de degré $g$ est alors constituée de l'ensemble des polynômes $P_{i,j}$ avec $i + j \leq g$ :

$$B_g = \{P_{0,0}, P_{0,1}, ..., P_{0,g}, P_{1,0}, P_{1,g-1}, ..., P_{g-1,0}, P_{g-1,1}, P_{g,0}\} \tag{3.7}$$

Le nombre de polynômes qui composent une base de degré $g$ est alors :

$$n_g = \frac{(g+1)(g+2)}{2} \tag{3.8}$$

### 3.2.2 Approximation d'un champ de déplacement

Pour exprimer le champ de déplacement $F$ par des combinaisons linéaires des différents polynômes $P_{i,j}$ de la base orthonormale $B_g$, nous projetons les fonctions $V^1(x_1, x_2)$ et $V^2(x_1, x_2)$ sur chaque polynôme $P_{i,j}$ de la base. Nous pouvons exprimer $F = (\tilde{V}^1(x_1, x_2), \tilde{V}^2(x_1, x_2))$ de la façon suivante :

$$\begin{cases} \tilde{V}^1(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}^1_{i,j} P_{i,j} \quad avec \quad \tilde{v}^1_{i,j} = \left\langle V^1(x_1, x_2) \mid P_{i,j} \right\rangle \\ \tilde{V}^2(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}^2_{i,j} P_{i,j} \quad avec \quad \tilde{v}^2_{i,j} = \left\langle V^2(x_1, x_2) \mid P_{i,j} \right\rangle \end{cases} \tag{3.9}$$

Les coefficients $\tilde{v}^1_{i,j}$ et $\tilde{v}^2_{i,j}$ sont calculés à partir du produit scalaire :

$$\begin{cases} \tilde{v}^1_{i,j} = \int \int_{\Omega} V^1(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \\ \tilde{v}^2_{i,j} = \int \int_{\Omega} V^2(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \end{cases} \tag{3.10}$$

avec $i + j \leq g$.

Connaissant les coefficients de projection $\tilde{v}^1_{i,j}$ et $\tilde{v}^2_{i,j}$, nous pouvons alors déterminer l'expression analytique du modèle grâce à l'équation 3.9. Finalement, à partir de cette expression analytique, nous pouvons reconstruire une approximation du champ original en évaluant les polynômes $\tilde{V}^1$ et $\tilde{V}^2$ obtenus, en tout point souhaité.

Par la suite, nous utilisons uniquement les polynômes de Legendre. La formule de récurrence 3.5 permettant de générer une base bidimensionnelle devient :

$$\begin{cases} P_{-1,j} = 0 \\ P_{i,-1} = 0 \\ P_{0,0} = 1 \\ P_{i+1,j} = \frac{2i+1}{i+1} x_1 P_{i,j} - \frac{i}{i+1} P_{i-1,j} \\ P_{i,j+1} = \frac{2j+1}{j+1} x_2 P_{i,j} - \frac{j}{j+1} P_{i,j-1} \end{cases} \tag{3.11}$$

Le choix des polynômes de Legendre se justifie pour plusieurs raisons. Tout d'abord, nous désirons effectuer une modélisation globale des champs de déplacement. La fonction de poids $\omega(x_1, x_2) = 1$ permet alors de donner la même importance à tous les vecteurs du champ. De plus, cette fonction de poids simplifie considérablement les équations permettant de calculer les coefficients de projection, ce qui diminue les temps de calcul. Un autre point concerne l'interprétation

physique du mouvement. En effet, pour certaines familles de polynômes, il est possible de donner une signification physique aux premiers polynômes de la base [3].

## 3.3 Tenseur d'orientation

Un tenseur d'orientation locale est un cas particulier d'une matrice, construit à partir des informations recueillies à partir d'une image. Ce type de tenseur a des propriétés particulières et contient de précieuses informations sur cette image [12].

Pour la définition de [13], le tenseur d'orientation est réel et symétrique, donc, peut être décomposé en utilisant le théorème spectrale de la façon suivante :

$$T = \sum_{i=1}^{n} \lambda_i T_i \tag{3.12}$$

où $\lambda_i$ sont les valeurs propres de $T$.

Si on projette $T_i$ sur un espace de dimension $m$, nous avons la décomposition suivante :

$$T_i = \sum_{s=1}^{m} e_s e_s^T \tag{3.13}$$

où $\{e_1,...,e_m\}$ est une base de $R^m$. Une décomposition intéressante du tenseur d'orientation $T$ est donnée par

$$T = \lambda_n T_n + \sum_{i=1}^{n-1} (\lambda_i - \lambda_{i+1}) T_i \tag{3.14}$$

où $\lambda_i$ sont les valeurs propres correspondant à chaque vecteur propre $e_i$. Cette décomposition-là est intéressante à cause de son interprétation géométrique. En effet, dans $\mathbb{R}^3$, le tenseur d'orientation $T$ décomposé en utilisant (3.14) peut être représenté par une lance (son orientation principale), un plat et un ballon.

$$T = (\lambda_1 - \lambda_2) T_1 + (\lambda_2 - \lambda_3) T_2 + \lambda_3 T_3. \tag{3.15}$$

Le tenseur de $R^3$ décomposé par (3.15), avec les valeurs propres $\lambda_1 \geq \lambda_2 \geq \lambda_3$, peut être interprété comme suit :

- $\lambda_1 >> \lambda_2 \approx \lambda_3$ correspond approximativement à un tenseur linéaire, avec la composante lance dominante.
- $\lambda_1 \approx \lambda_2 >> \lambda_3$ correspond approximativement à un tenseur plat, avec la composante plat dominante.
- $\lambda_1 \approx \lambda_2 \approx \lambda_3$ correspond approximativement à un tenseur isotropique, avec la composante ballon dominante, et pas d'orientation dominante.

Nous pouvons créer ce type de tenseur à partir d'un vecteur en utilisant sa transposée de la façon suivante :

$$T = vv^T \tag{3.16}$$

où $T$ est le tenseur crée à partir du produit entre le vecteur $v$ et sa transposée. Donc, le tenseur $T$ est une matrice $n$ x $n$, où $n$ est la taille du vecteur $v$.

## 3.4 Machines à Vecteurs de Support (SVM)

Le SVM (*support vectors machines* ou machines à vecteurs de support) est une méthode de classification à hyperplan séparateur. Le SVM se distingue des autres méthodes à hyperplan par le critère qu'elle optimise, dont la conséquence est le choix de l'hyperplan qui maximise la marge.

Dans la Figure 3.3, on détermine un hyperplan $H$ qui sépare les deux ensembles de points. Les points les plus proche, qui seuls sont utilisé pour la détermination de l'hyperplan, son appelés vecteurs de support. L'hyperplan séparateur optimal est celui qui maximise la marge.

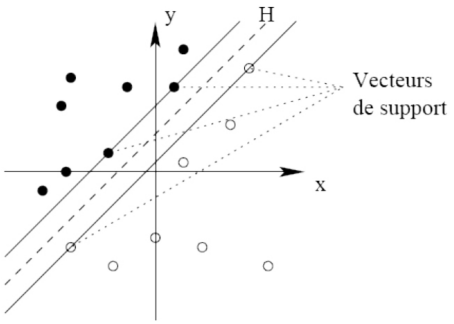



FIGURE 3.3 – Hyperplan qui sépare les données et les vecteurs de support

Cette méthode est compatible avec l'utilisation de fonction dites noyaux (ou *kernel*) qui permettent une séparation non linéaire des données. Quelques fonctions noyaux classiques sont :

– **Linéaire** :

$$k(x, y) = \langle x, y \rangle$$

– **Polynomiale de degré d** :

$$k(x, y) = \langle x, y \rangle^d$$

– **Gaussien avec une distance Euclidienne** (Gaussien $L^2$) :

$$k(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}}$$

– **Gaussien avec une distance du $\chi^2$** (Gaussien $\chi^2$) :

$$k(x, y) = e^{-\frac{d(x,y)^2}{2\sigma^2}}$$

avec $d$ une distance au sens du $\chi^2$ :

$$d(x,y) = \sum_{r=1}^{p} \frac{(x_r - y_r)^2}{x_r + y_r}$$

– **Triangulaire** :

$$k(x,y) = 1 - \frac{1}{\sigma^2}||x - y||$$

L'avantage des méthodes par fonction noyaux est qu'on peut travailler dans les espaces non vectoriels, par exemple, les espaces de sac de descripteurs. Étant $B_i = b_{r_i i}$ le sac de descripteurs des régions $r_i$ d'une image $i$, un exemple d'usage de la fonction noyau par le problème de classification est donné par la façon suivante :

$$K(B_i, B_j) = \frac{1}{2|B_i|} \sum_r max_s\{k(b_{r_i}, b_{r_j})\} + \frac{1}{2|B_j|} \sum_s max_r\{k(b_{r_i}, b_{r_j})\} \quad (3.17)$$

où $K(B_i, B_j)$ est la distance entre les ensembles de descripteurs $B_i$ et $B_j$, et $k$ est une fonction noyau gaussien $L^2$.

# Chapitre 4

# Proposition d'un descripteur tensoriel

Tout d'abord, nous proposons d'utiliser les coefficients $\tilde{v}^1_{i,j}$ et $\tilde{v}^2_{i,j}$ (3.10) pour créer un vecteur pour chaque *frame* :

$$vc = [\tilde{v}^1_{0,0}, ..., \tilde{v}^1_{g,0}, \tilde{v}^2_{0,0}, ..., \tilde{v}^2_{g,0}] \in \mathbb{R}^m \tag{4.1}$$

À partir de ce vecteur-là, nous créons un tenseur d'orientation pour chaque *frame* $f$ :

$$T_f = vc\ vc^T \tag{4.2}$$

Le descripteur est formé par la somme des tenseurs significatifs des *frames* d'une vidéo (Eq. 4.3) :

$$\bar{D} = \sum_{f=a}^{b} T_f \tag{4.3}$$

où $[a, b]$ est l'intervalle des *frames* qui contiennent le mouvement le plus représentatif de la vidéo. Ainsi, on fait une combinaison m-dimensionnelle qui capture la dispersion des tenseurs dans cet espace. Il est possible d'utiliser seulement un morceaux de vidéo pour composer le descripteur, c'est-à-dire, nous pouvons choisir la partie la plus représentative du mouvement.

La proposition de ce travail est d'utiliser le tenseur résultant pour exprimer le mouvement moyen de plusieurs *frames* consécutives. C'est possible car l'espace propre du tenseur final capture des informations importantes comme, par exemple, des anisotropies qui doivent être exploitées. Soit $\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, les valeurs propres d'un tenseur $T$. Le plus la valeur $\lambda_1 - \lambda_2$ est grande, mieux le tenseur exprime un mouvement moyen donné par le vecteur propre principal. C'est à dire, des tenseurs anisotropes sont nécessaires pour bien capturer le changement des champs de déplacement.

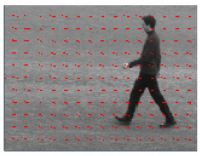
Avec deux tenseurs $A$ et $B$, on peut définir leur distance en utilisant la norme de Frobenius :

$$||A - B|| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} |A_{ij} - B_{ij}|^2} \tag{4.4}$$

## 4.1 Application du descripteur pour la recherche des vidéos

Dès qu'on calcul le flot optique, nous avons les champs de déplacement pour chaque frame de vidéo. À partir de ces champs, on fait la modélisation par base de polynômes et on rencontre $n_g$ coefficients pour le polynôme en $x_1$ et pour le polynôme en $x_2$. Ainsi, nous avons $2 * n_g$ coefficients pour chaque frame du vidéo.

La Figure 4.1 montre un exemple de comment un tenseur est formé à partir des coefficients d'une base de degré 1. Pour cette base nous avons 6 coefficients pour chaque frame (3 pour le polynôme en $x_1$ et 3 le polynôme en $x_2$), donc le tenseur formé est une matrice 6x6.



Polynômes genérés

$$V^1 = -0.467823 - 0.366877x_1 - 0.061723x_2$$
$$V^2 = 0.279198 + 0.224533x_1 - 0.06561x_2$$

Tenseur genéré

$$T = \begin{matrix} 0.218858 & 0.171634 & 0.028875 & -0.130615 & -0.105042 & 0.030697 \\ 0.171634 & 0.134599 & 0.022645 & -0.102431 & -0.082376 & 0.024073 \\ 0.028875 & 0.022645 & 0.003810 & -0.017233 & -0.013859 & 0.004050 \\ -0.130615 & -0.102431 & -0.017233 & 0.077951 & 0.062689 & -0.018320 \\ -0.105042 & -0.082376 & -0.013859 & 0.062689 & 0.050415 & -0.014733 \\ 0.030697 & 0.024073 & 0.004050 & -0.018320 & -0.014733 & 0.004306 \end{matrix}$$

FIGURE 4.1 – Création du tenseur à partir des coefficients $\tilde{v}^1_{i,j}$ et $\tilde{v}^2_{i,j}$

Le descripteur est donc composé par la somme des tenseurs des *frames* les plus représentatives de la vidéo. Pour pouvoir comparer les descripteurs de différentes vidéos, ils doivent être normalisés, car on ne sait pas combien de *frames* nous avons utilisé pour les composer. Nous pouvons faire une compa-

raison entre les descripteurs en utilisant la distance calculée par la norme de Frobenius (Eq. 4.4).

Une partie importante dans la composition du descripteur est le choix des parties les plus représentatives. Il faut calculer l'anisotropie pour que les tenseurs qui ne sont pas représentatifs du mouvement ne fassent pas partie du calcul. L'anisotropie $\alpha$ d'un tenseur $T \in R^m$ est calculée de la façon suivante :

$$\alpha_T = \frac{\sum_{i=1}^{m-1}(\lambda_i - \lambda_{i+1})^2}{2\sum_{i=1}^{m}\lambda_i^2} \tag{4.5}$$

Au lieu d'avoir un seul descripteur par vidéo, nous pouvons aussi composer plusieurs descripteurs. De cette façon nous avons un ensemble de caractéristiques. Comme cela, au lieu d'avoir une moyenne de tout le mouvement de la vidéo, nous avons la moyenne de chaque morceaux du mouvement de la vidéo.

# Chapitre 5

# Analyse des Résultats

Le descripteur $D$ de base de degré 1, décrit par l'équation 4.3, a été crée à partir de tous les *frames* d'un vidéo du mouvement *walking* et est donné par la matrice 6x6 suivante :

$$D = \begin{matrix} 0.749298 & 0.055895 & -0.049645 & -0.052788 & 0.012378 & -0.016854 \\ 0.055895 & 0.642323 & -0.042963 & -0.008069 & -0.042564 & 0.007108 \\ -0.049645 & -0.042963 & 0.015371 & 0.001108 & 0.001498 & 0.001617 \\ -0.052788 & -0.008069 & 0.001108 & 0.015298 & 0.001644 & -0.004647 \\ 0.012378 & -0.042564 & 0.001498 & 0.001644 & 0.014899 & -0.003446 \\ -0.016854 & 0.007108 & 0.001617 & -0.004647 & -0.003446 & 0.004976 \end{matrix}$$

Dans la Figure 5.1 on peut voir les différents comportements du descripteur $D$ du mouvement *walking* comparé morceaux par morceaux (*blocks* de 16 *frames*) avec trois types de vidéo : "sa propre vidéo", une vidéo différente du mouvement *walking* et une vidéo du mouvement *boxing*. Il est possible de conclure que pour les mêmes mouvements nous avons les distributions de distances semblables. De plus, nous avons la même distribution quand on augmente la taille de la base (Figure 5.2). Il est importante de remarquer que les parties plus stables du graphique correspondent aux parties de la vidéo qui n'ont pas de mouvement.

Le tableau 5.1 montre les distances moyennes et le tableau 5.2 montre les variances trouvées entre les ensembles de vidéos pour une base de degré 1.

À partir des ces distances moyennes et ses variances, nous pouvons voir que le descripteur peut séparer les mouvements *boxing*, *hand waving*, *hand clapping*, par contre, les mouvements *walking*, *jogging* et *running* sont considérés proches. En effet, visiblement ces mouvements sont très proches et une base de polynômes de degré 1 n'est pas suffisante pour bien les décrire. Pour résoudre ce problème, nous avons fait les mêmes calculs pour une base de degré 9 (Tableaux 5.3 et 5.4).

Il est possible de conclure que les résultats sont améliorés et on commence à avoir une distance plus grande entre les ensembles *walking*, *jogging* et *running*, mais ces distances sont encore très proches. Par contre, ces résultats montrent

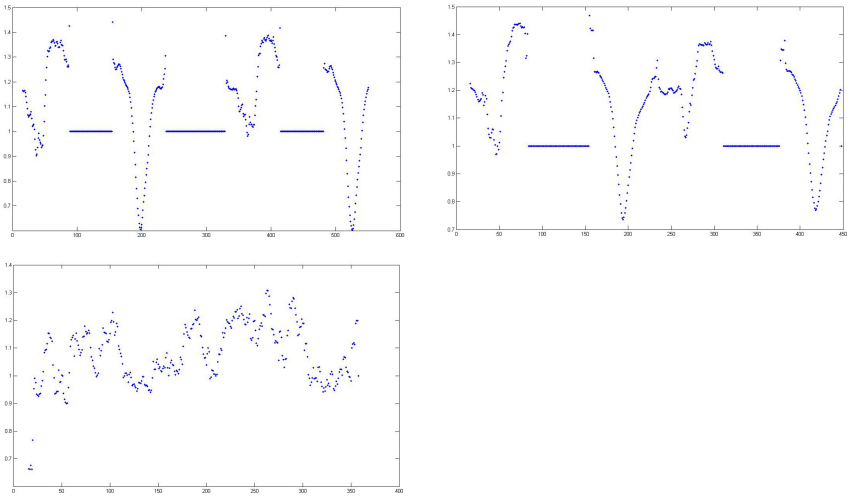


FIGURE 5.1 – Descripteur $D$ de base de degré 1 comparé morceaux par morceaux avec sa vidéo, l'autre vidéo du mouvement *walking* et un vidéo du mouvement *boxing*

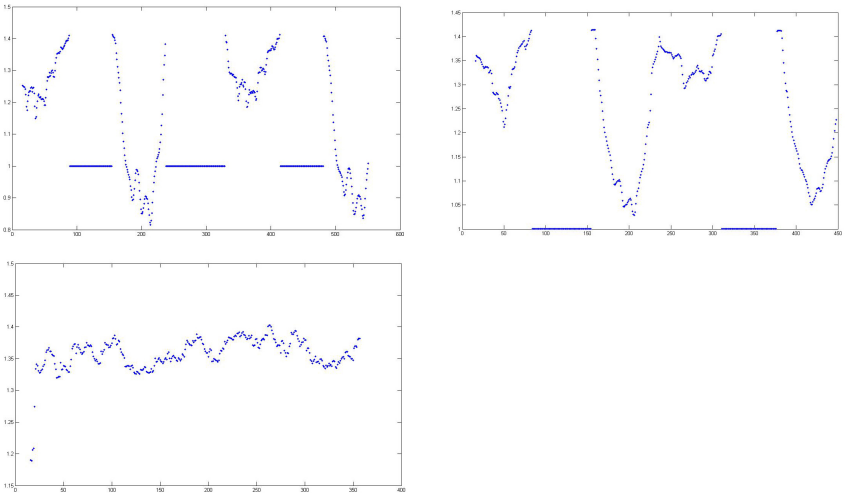


FIGURE 5.2 – Descripteur $D$ de base de degré 9 comparé morceaux par morceaux avec sa vidéo, l'autre vidéo du mouvement *walking* et un vidéo du mouvement *boxing*

| | Walk | Box | HWav | HClap | Jog | Run |
|---|---|---|---|---|---|---|
| Walk | 0,509 | 1,139 | 1,228 | 1,032 | 0,499 | 0,523 |
| Box | | 0,696 | 0,737 | 0,832 | 1,121 | 1,106 |
| HWav | | | 0,524 | 0,835 | 1,216 | 1,195 |
| HClap | | | | 0,733 | 1,011 | 0,982 |
| Jog | | | | | 0,476 | 0,510 |
| Run | | | | | | 0,503 |

TABLE 5.1 – Distances moyennes entre les ensembles avec une base de degré 1

| | Walk | Box | HWav | HClap | Jog | Run |
|---|---|---|---|---|---|---|
| Walk | 0,062 | 0,030 | 0,008 | 0,029 | 0,053 | 0,046 |
| Box | | 0,079 | 0,050 | 0,050 | 0,031 | 0,029 |
| HWav | | | 0,058 | 0,051 | 0,009 | 0,010 |
| HClap | | | | 0,068 | 0,029 | 0,027 |
| Jog | | | | | 0,056 | 0,048 |
| Run | | | | | | 0,052 |

TABLE 5.2 – Variances entre les ensembles avec une base de degré 1

| | Walk | Box | HWav | HClap | Jog | Run |
|---|---|---|---|---|---|---|
| Walk | 1,067 | 1,307 | 1,313 | 1,283 | 1,165 | 1,225 |
| Box | | 0,955 | 1,223 | 1,225 | 1,311 | 1,328 |
| HWav | | | 1,008 | 1,193 | 1,309 | 1,320 |
| HClap | | | | 1,063 | 1,292 | 1,307 |
| Jog | | | | | 1,159 | 1,219 |
| Run | | | | | | 1,198 |

TABLE 5.3 – Distances moyennes entre les ensembles avec une base de degré 9

| | Walk | Box | HWav | HClap | Jog | Run |
|---|---|---|---|---|---|---|
| Walk | 0,067 | 0,008 | 0,005 | 0,011 | 0,024 | 0,016 |
| Box | | 0,083 | 0,011 | 0,023 | 0,006 | 0,005 |
| HWav | | | 0,066 | 0,020 | 0,004 | 0,003 |
| HClap | | | | 0,079 | 0,006 | 0,005 |
| Jog | | | | | 0,059 | 0,017 |
| Run | | | | | | 0,059 |

TABLE 5.4 – Variances entre les ensembles avec une base de degré 9

| Degré de la base | Précision |
|---|---|
| 1 | 70% |
| 2 | 77,7% |
| 3 | 76% |
| 4 | 77,5% |
| **5** | **78,2%** |
| 6 | 75% |
| 7 | 73,8% |
| 8 | 72,6% |
| 9 | 68,2% |

TABLE 5.5 – Précisions

| | Box | HWav | HClap | Jog | Run | Walk |
|---|---|---|---|---|---|---|
| Box | **93,00%** | *0,69%* | *1,39%* | *0,69%* | *4,19%* | 0,0% |
| HWav | *5,55%* | **82,63%** | *11,80%* | 0,0% | 0,0% | 0,0% |
| HClap | *16,66%* | *2,08%* | **81,25%** | 0,0% | 0,0% | 0,0% |
| Jog | 0,0% | 0,0% | 0,0% | **59,02%** | *19,44%* | *21,52%* |
| Run | 0,0% | 0,0% | 0,0% | *17,36%* | **77,77%** | *4,86%* |
| Walk | 0,0% | 0,0% | 0,0% | *9,72%* | *4,16%* | **86,11%** |

TABLE 5.6 – Matrice de confusion pour la base de degré 5

que les distances sont séparables et que nous pouvons tester le descripteur avec un classifieur SVM.

Ainsi, nous avons testé le descripteur avec un classifieur SVM. Le Tableau 5.5 montre les précisions trouvées pour les bases de degré 1 à 9.

Il est intéressant de remarquer que le plus haut le degré de la base, le meilleur est la modélisation du mouvement, par contre il devient plus difficile à classifier. En effet, le plus haut le degré de la base, plus de coefficients nous avons et cela peut déranger le classifieur. Par exemple, le descripteur composé par les coefficients de la base de degré 5 a 1784 éléments tandis que le descripteur de la base de degré 9 a 12100 éléments. En fonction de cela, plus on augmente le degré de la base, le descripteur devient de plus en plus spécifique et très différent de tous les autres descripteurs.

Nous pouvons voir que le meilleur résultat a été obtenu pour la base de degré 5 $(78,2\%)$. Pour essayer de l'améliorer, nous avons changé les paramètres du classifieur SVM. En effet, avec le changement de la fonction noyau gaussien pour une fonction noyau triangulaire de $\sigma = 1.2$, nous avons trouvée une précision de $79,96\%$. La matrice de confusion est donnée par le tableau 5.6.

Avec la matrice de confusion, nous pouvons voir que le mélange le est entre l'ensemble *jogging* et les ensembles *walking* et *running*.

Selon ces résultats, la précision maximum que nous avons trouvée est de $79,96\%$ qui est encore pas bonne par rapport aux précisions trouvées dans l'état

| Approche | Précision |
|---|---|
| [10] | 94,0% |
| [11] | 91,2% |
| **Notre approche** | **79,96%** |
| [5] | 71,7% |

TABLE 5.7 – Précisions

| Degré de la base | Précision |
|---|---|
| 1 | 61,32% |
| 2 | 69,20% |
| 3 | 71,40% |
| **5** | **74,41%** |

TABLE 5.8 – Précisions

de l'art (Tableau 5.7). Par contre, les descripteurs trouvés dans [10], [11] et [5] sont composé par plus d'une caractéristiques globales et des caractéristiques locales. Avec cela, nous pouvons conclure que parmi les descripteurs, notre descripteur étant global a donné une précision très bonne.

Une idée pour améliorer les résultats est d'avoir un ensemble de tenseurs (*bag of tensors*). Comme cela, au lieu d'avoir une moyenne de tout le mouvement de la vidéo, nous avons une moyenne de chaque morceaux de mouvement de la vidéo.

Nous avons donc composé plusieurs tenseurs à partir des morceaux de 16 *frames* de la vidéo. La précision trouvée par le classifieur SVM pour une base de degré 1 a été de $61,32\%$. Ainsi, au-delà d'être une approche plus lourde, elle a donné de moins bons résultats pour l'instant. Même quand on augmente la taille de la base, la précision est encore petite par rapport à celle avec un seul descripteur (Tableau 5.8).

L'avantage de cette approche est que la précision paraît augmenter avec le degré de la base, par contre le plus grande est le degré de la base, le plus lente devient la classification.

# Chapitre 6

# Conclusion et Perspectives

Dans ce travail, nous avons proposé un descripteur global pour les actions humaines de la base KTH basé sur le tenseur d'orientation. Ce tenseur d'orientation est composé à partir des coefficients des polynômes de Legendre calculés pour chaque *frame* d'une vidéo.

Pour tester le descripteur créé nous avons classifié la base KTH avec un classifieur SVM. Nous avons trouvé la meilleure précision égale à 79.96% avec une base de degré 5. Le plus grand problème de classification a été entre l'ensemble *jogging* et les ensembles *running* et *walking*. En effet, cela est un problème qui apparaît dans plusieurs articles.

Nous avons vu que ce descripteur donne une précision encore loin des meilleures précisions trouvées dans l'état de l'art, par contre, étant un descripteur seulement global, nous pouvons conclure que c'est un descripteur prometteur. En fonction de cela, l'étude des caractéristiques spectrales des tenseurs créés paraît intéressante.

Nous avons aussi testé l'idée de faire un sac de descripteurs, mais les résultats préliminaires ne sont pas meilleurs.

En plus, il est possible de combiner ce descripteur avec autres descripteurs composés par d'autres caractéristiques, comme nous pouvons voir dans l'état de l'art.

# Bibliographie


[1] F.B. Lauze, P. Kornprobst, C. Lenglet, R. Deriche, and M. Nielsen, "Sur quelques méthodes de calcul de flot optique à partir du tenseur de structure : Synthèse et contribution," 2004.

[2] B. Augereau, B. Tremblais, and C. Fernandez-Maloigne, "Vectorial computation of the optical flow in color image sequences.," in *Thirteenth Color Imaging Conference*, November 2005, pp. 130–134.

[3] Martin Druon, *Modélisation du mouvement par polynômes orthogonaux : application à l'étude d'écoulements fluides*, Ph.D. thesis, Université de Poitiers, 02 2009.

[4] M. Druon, B. Tremblais, and B. Augereau, "Modélisation de champs de vecteurs par bases de polynômes : application à l'analyse de la posture d'utilisateurs devant un écran d'ordinateur, via une webcam.," in *COmpression et REprésentation des Signaux Audiovisuels (CORESA)*, Caen, France, Novembre 2006, pp. 290–295.

[5] Christian Schüldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions : A local svm approach," in *In Proc. ICPR*, 2004, pp. 32–36.

[6] Riemenschneider Hayko, Donoser Michael, and Bischof Horst, "Bag of optical flow volumes for image sequence recognition," 2009, Proceedings of British Machine Vision Conference (BMVC), 2009.

[7] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," 2007, pp. 214–223, Proceedings of Symposium of the German Association for Pattern Recognition (DAGM.

[8] Xinbo Gao, Yimin Yang, Dacheng Tao, and Xuelong Li, "Discriminative optical flow tensor for video semantic analysis," *Comput. Vis. Image Underst.*, vol. 113, no. 3, pp. 372–383, 2009.

[9] Xinbo Gao, Xuelong Li, Jun Feng, and Dacheng Tao, "Shot-based video retrieval with optical flow tensor and hmms," *Pattern Recogn. Lett.*, vol. 30, no. 2, pp. 140–147, 2009.

[10] Nazli Ikizler, R. Gokberk Cinbis, and Pinar Duygulu, "Human action recognition with line and flow histograms," in *In Proc. ICPR*, 2008.

[11] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra, "Recognizing human actions by fusing spatio-temporal

appearance and motion descriptors," in *ICIP'09 : Proceedings of the 16th IEEE international conference on Image processing*, Piscataway, NJ, USA, 2009, pp. 3533–3536, IEEE Press.

[12] H. Knutsson, "Representing local structure using tensors," in *The 6th Scandinavian Conference on Image Analysis*, Oulu, Finland, June 1989, pp. 244–251, Report LiTH–ISY–I–1019, Computer Vision Laboratory, Linköping University, Sweden, 1989.

[13] C.-F. Westin, *A Tensor Framework for Multidimensional Signal Processing*, Ph.D. thesis, Linköping University, Sweden, S-581 83 Linköping, Sweden, 1994, Dissertation No 348, ISBN 91-7871-421-4.