# Indexation des Bases Vidéos à l'aide d'une Modélisation du Flot Optique par Bases de Polynômes

Romain Negrel, Virgínia Fernandes Mota, Philippe-Henri Gosselin, Marcelo Bernardes Vieira, Frederic Precioso

# Indexation des Bases Vidéos à l'aide d'une Modélisation du Flot Optique par Bases de Polynômes

R. Negrel[1] V. Fernandes[2] P.H. Gosselin[1] M. Vieira[2] F. Precioso[3]

[1] ETIS, CNRS, ENSEA, Université Cergy-Pontoise F-95000 Cergy, France
[2] Universidade Federal de Juiz de Fora, Brazil
[3] I3S - UMR 6070 UNS/CNRS, Université Nice-Sophia Antipolis, France

## Résumé

*La classification d'action dans les vidéos est un problème qui a pris de plus en plus d'importance ces dernières années dans la communauté de reconnaissance des formes. Nous proposons une méthode basée sur la modélisation du flot optique par une base de polynômes et une représentation innovante en sac de sacs de descripteurs de flot optique. Nous utilisons une classification par SVM et les bases vidéo KTH et Hollywood2 pour évaluer la performance de notre méthode sur la classification d'actions humaines dans les vidéos. Nos résultats démontrent que ces performances sont au moins comparables aux travaux les plus récents avec une approche plus simple et plus rapide.*

## Mots Clef

Indexation, sac de descripteurs de flot optique, reconnaissance d'actions humaines.

## Abstract

*Action classification in videos is a problem that has become increasingly important in recent years in the pattern recognition community. We propose a method based on optical flow modeling using polynomial basis and a new representation based on bags of bags of optical flow descriptors. We use a SVM classification with the KTH and Hollywood2 databases for evaluating the performance of our method over human action classification on video. Our results show that this performance is comparable to recent works with a much simpler and faster approach.*

## Keywords

Indexation, bag of optical flow descriptors, human action recognition.

# 1 Introduction

La reconnaissance automatique des actions humaines dans les vidéos est une des problématiques les plus complexes de la vision par ordinateur. Avec l'accroissement de la taille des bases vidéos tant personnelles que professionnelles, ces méthodes deviennent des points clefs des applications d'indexation et de classification vidéo.

La problématique de la reconnaissance d'action suscite l'intérêt depuis les premières bases vidéos et a énormément progressé ces dernières années. Les premiers travaux dans ce domaine consistaient à positionner, à la fois temporellement et spatialement, les actions dans la vidéo, alors que de nos jours on recherche à identifier ces actions.

Les méthodes récentes d'identification d'actions donnent de très bons résultats dans les bases vidéos d'actions en milieu contrôlé, par exemple dans les bases KTH [10] et Weizman [4] . Mais pour le moment, aucune méthode n'a fourni de résultats réellement satisfaisants dans les bases de vidéos réelles (films, vidéo surveillance, ...), comme cela a été montré dans [2] et [6]. Il est donc essentiel d'éprouver les nouvelles méthodes développées sur des bases de vidéos réelles.

Dans ce travail nous proposons un nouveau schéma d'indexation pour la reconnaissance d'action dans les vidéos. Pour la description du mouvement, nous proposons d'utiliser une modélisation du flot proposée par Druon dans [3]. Cette technique modélise le flot optique entre deux frames successives de la vidéo à l'aide d'une base de polynômes orthogonaux. Puis nous décrivons une méthode originale pour le calcul de signatures et de métriques compatibles avec la classification par SVM. Cette méthode s'appuie sur le formalisme des fonctions noyaux, et sur les méthodes basées sur le principe du "coding-pooling" [13].

Enfin, nous utilisons la base vidéo KTH pour évaluer de la méthode proposée sur des actions humaines en milieu contrôlé. Puis nous utilisons la base Hollywood2 [9] pour effectuer un test de la méthode proposée sur des actions humaines en milieu réel.

# 2 Description vectorielle du flot optique

Dans cette section, nous présentons la méthode de description vectorielle du flot optique proposée par Druon [3]. Cette méthode s'appuie tout d'abord sur une extraction du flot optique sous la forme d'un champ de vecteur. Cette extraction est effectuée par le biais de la methode de Lucas et Kanade [7]. Puis le champ de vecteur est modélisé à l'aide d'une projection sur une base de polynômes, dont les coef-

ficients forment la description vectorielle du flot entre deux frames successives.

## 2.1 Extraction du mouvement apparent

L'estimation du mouvement consiste à mesurer la projection 2D dans le plan de l'image d'un mouvement réel 3D. Le flot optique est défini comme le champ de vitesse décrivant le mouvement apparent des motifs d'intensité de l'image sous l'hypothèse d'illumination constante :

$$\frac{\partial I}{\partial x_1} v_{x_1} + \frac{\partial I}{\partial x_2} v_{x_2} + \frac{\partial I}{\partial t} = 0 \tag{1}$$

avec $(v_{x_1}, v_{x_2}) \in \mathbb{R}^2$ les composantes horizontale et verticale du flot optique et

$$\begin{array}{rl} I: & \mathcal{V} \subset \mathbb{R}^3 \to [0,1] \\ & (x_1, x_2, t) \to I(x_1, x_2, t) \end{array} \tag{2}$$

la fonction de luminance de la vidéo.
La méthode de Lucas et Kanade [7] est une approche locale différentielle pour l'extraction du flot optique. Elle fait l'hypothèse que le flot optique est localement constant sur un voisinage spatial. Cette méthode détermine le déplacement d'un pixel $\mathbf{x} = (x_1, x_2, t) \in \mathcal{V}$ à partir de l'information des pixels voisins dans une fenêtre $W(\mathbf{x}) \subset \mathcal{V}$ (Eq. (3)). Le flot $\vec{v}(\mathbf{x})$ est obtenu avec la minimisation de cette énergie :

$$\min_{\vec{v}(\mathbf{x})} \int_{W(\mathbf{x})} \vec{v}(\mathbf{x})^\top (\nabla I(x'))(\nabla I(x'))^\top \vec{v}(\mathbf{x}) dx'. \tag{3}$$

Au lieu d'avoir un simple moyennage du voisinage de $\mathbf{x}$, il est possible de pondérer l'équation par une fonction $h(x)$ qui est, généralement, un noyau gaussien de moyenne nulle et d'écart type $s_x$

$$h(x) = \frac{1}{2\pi s_x^2} e^{-\frac{x^\top x}{2 s_x^2}}. \tag{4}$$

Si nous supposons que :

$$\langle f \rangle (\mathbf{x}) = \int_{W(\mathbf{x})} h(\mathbf{x} - x') f(x', t) dx' \tag{5}$$

et que

$$S = \left\langle (\nabla I)(\nabla I)^\top \right\rangle \tag{6}$$

est la définition du tenseur de structure, alors, le problème de l'Eq. (3) peut être réécrit comme :

$$\min_{\vec{v}=(v_{x_1}, v_{x_2}, 1)} \vec{v}^\top S \vec{v}. \tag{7}$$

La minimisation de l'énergie est un problème de moindres carrés pondérés et la solution revient à considérer le système d'equations suivant :

$$A \begin{pmatrix} v_{x_1} \\ v_{x_2} \end{pmatrix} = b \tag{8}$$

où

$$A = \begin{pmatrix} \langle I_{x_1}^2 \rangle & \langle I_{x_1} I_{x_2} \rangle \\ \langle I_{x_1} I_{x_2} \rangle & \langle I_{x_2}^2 \rangle \end{pmatrix} \text{ et } b = \begin{pmatrix} \langle I_{x_1} I_t \rangle \\ \langle I_{x_2} I_t \rangle \end{pmatrix}. \tag{9}$$

La méthode de Lucas et Kanade permet de corriger localement les problèmes d'ouvertures, mais il peuvent persister si des parties de la vidéo (plus grande que le voisins $W(\mathbf{x})$) ont une structure linéaire.

## 2.2 Modélisation par polynôme de Legendre

L'idée de la modélisation par polynômes orthogonaux est d'approximer des fonctions réelles par des combinaisons linéaires de fonctions polynomiales [3]. Druon et al. approximent ainsi le flot optique par des polynômes d'une base orthogonale, par exemple celle de Legendre. L'orthogonalité permet de ne pas introduire de redondance dans la décomposition. On peut définir le flot optique $F$ à un instant $t$ de la vidéo comme suit :

$$\begin{array}{rl} \vec{v}: & \mathcal{I}_t \to \mathbb{R}^2 \\ & (x_1, x_2) \to (V^1(x_1, x_2), V^2(x_1, x_2)) \end{array} \tag{10}$$

où $\mathcal{I}_t = \{(x_1, x_2, t_\mathcal{V}) \in \mathcal{V} \mid t_\mathcal{V} = t\}$ et, $V^1(x_1, x_2)$ et $V^2(x_1, x_2)$ sont deux applications correspondant respectivement aux déplacements horizontaux et verticaux au point de coordonnées $(x_1, x_2)$.
Nous utilisons des polynômes définis dans $\mathbb{R}^2$ de la façon suivante :

$$P(x_1, x_2)_{K,L} = \sum_{k=0}^{K} \sum_{l=0}^{L} c_{k,l} (x_1)^k (x_2)^l \tag{11}$$

avec $K \in \mathbb{N}^+$ le degré maximal de $x_1$, $L \in \mathbb{N}^+$ le degré maximal de $x_2$ et $c_{k,l}$ l'ensemble des coefficients réels du polynôme. Le dégre du polynôme est $K+L$ [3]. Le produit scalaire dans les bases de polynômes bidimensionnelles est définie par :

$$\langle f | g \rangle = \iint_\Omega f(x_1, x_2) g(x_1, x_2) \omega(x_1, x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \tag{12}$$

avec $\omega(x_1, x_2)$ la fonction de poids du produit scalaire.

**Polynôme de Legendre.** Les polynômes de Legendre sont des solutions de l'équation différentielle de Legendre, et constituent un exemple de base de polynômes orthogonaux.
Ils peuvent être construits par la formule de récurrence suivante :

$$\begin{cases} P_{-1,j} = 0 \\ P_{i,-1} = 0 \\ P_{0,0} = 1 \\ P_{i+1,j} = \frac{2i+1}{i+1} x_1 P_{i,j} - \frac{i}{i+1} P_{i-1,j} \\ P_{i,j+1} = \frac{2j+1}{j+1} x_2 P_{i,j} - \frac{j}{j+1} P_{i,j-1} \end{cases}. \tag{13}$$

Une base bidimensionnelle de degré $d$ peut être composée par les polynômes de Legendre $\{P_{i,j}\}$ avec $i + j \leq d$. Le

nombre de polynômes qui composent la base de degré $d$ est :

$$n_d = \frac{(d+1)(d+2)}{2}. \quad (14)$$

Le domaine d'orthogonalité des polynômes de Legendre bidimensionnels est $\Omega \in [-1,1]^2$. Il est important de remarquer que la fonction de poids du produit scalaire $\omega(x_1, x_2)$ vaut 1, cela rend le calcul du produit scalaire plus simple et plus rapide, contrérement à de nombreuses bases de polynômes orthogonales qui ont des fonctions de poids qui demande un temps de calcul non négligeable. Le produit scalaire dans la base de polynômes de Legendre peut être calculé en évaluant que deux fonctions au lieu de trois.

**Projection dans la base.** La modélisation du flot optique est générée à partir de la projection de $V^1(x_1, x_2)$ et $V^2(x_1, x_2)$ sur chaque polynôme $P_{i,j}$ de la base orthogonale de degré d. L'approximation du flot optique $v = (\tilde{V}^1(x_1, x_2), \tilde{V}^2(x_1, x_2))$ peut être exprimée comme :

$$\begin{cases} \tilde{V}^1(x_1, x_2) = \sum_{i=0}^{d} \sum_{j=0}^{d-1} \tilde{v}^1_{i,j} P_{i,j} \\ \tilde{V}^2(x_1, x_2) = \sum_{i=0}^{d} \sum_{j=0}^{d-1} \tilde{v}^2_{i,j} P_{i,j} \end{cases} \quad (15)$$

avec

$$\begin{cases} \tilde{v}^1_{i,j} = \langle V^1 | P_{i,j} \rangle \\ \tilde{v}^2_{i,j} = \langle V^2 | P_{i,j} \rangle \end{cases}. \quad (16)$$

Ce processus de modélisation nous permet de décrire chaque paire $r$ de frames successives de chaque vidéo $s$ sous la forme d'un vecteur $\mathbf{b}_{rs} \in \mathbb{R}^{2n_d}$. Ce vecteur $\mathbf{b}_{rs}$ contient l'ensemble des coefficients $\tilde{v}^1_{i,j}$ et $\tilde{v}^2_{i,j}$ issus de la projection du champ de vecteurs sur la base polynomiale choisie.

# 3 Reconnaissance d'action

## 3.1 Approche proposée

Nous proposons d'utiliser le descripteur présenté dans la section précédente pour la classification d'actions dans les vidéos. Notre méthode se base sur des tubes vidéos, où un tube vidéo est un sous-ensemble d'une vidéo (de forme parallélépipédique rectangle dans notre cas). Chaque tube $i$ est décrit par un sac $B_i \in \mathcal{B}$ de descripteurs $\mathbf{b}_{ri}$, où chaque descripteur correspond à une paire de frames successives, notée $r$.

Puis, nous proposons de munir l'espace $\mathcal{B}$ avec une métrique compatible avec les méthodes de classification par hyperplan telles que les Séparateurs à Vaste Marge (SVM) [12].

Les données à classifier doivent être décrites dans un espace hilbertien $\mathcal{H}$. Dans le but de créer cet espace, nous avons alors choisi d'utiliser la méthode par noyau. Cette méthode propose de transformer l'espace initial $\mathcal{B}$ via une fonction $\phi : \mathcal{B} \to \mathcal{H}$, puis on travaille sur la métrique dans l'espace $\mathcal{H}$, appelée fonction noyau :

$$k(B_i, B_j) = \langle \phi(B_i) | \phi(B_j) \rangle. \quad (17)$$

Dans notre cas particulier où l'espace initial est un ensemble de sacs, nous avons besoin d'utiliser une fonction noyau dite sur sacs. Parmi celles-ci, nous avons basé notre modèle sur la fonction proposée par Lyu [8], avec $p \geq 1$ et $k(.,.)$ une fonction noyau entre vecteurs :

$$K(B_i, B_j) = \sum_r \sum_s k(\mathbf{b}_{ri}, \mathbf{b}_{sj})^p. \quad (18)$$

Pour $p = 2$ et $k(\cdot,\cdot) = \langle \cdot | \cdot \rangle$, on peut expliciter $\phi$ :

$$\begin{aligned} K(B_i, B_j) &= \sum_r \sum_s \langle \mathbf{b}_{ri} | \mathbf{b}_{sj} \rangle^2 \\ &= \sum_r \sum_s (\mathbf{b}_{ri}^\top \mathbf{b}_{sj})^2 \\ &= \text{trace}\left( \sum_r \sum_s (\mathbf{b}_{ri} \mathbf{b}_{ri}^\top)^\top \mathbf{b}_{sj} \mathbf{b}_{sj}^\top \right) \\ &= \text{vect}\left( \sum_r \mathbf{b}_{ri} \mathbf{b}_{ri}^\top \right)^\top \text{vect}\left( \sum_s \mathbf{b}_{sj} \mathbf{b}_{sj}^\top \right) \end{aligned} \quad (19)$$

avec $\text{vect}(\cdot)$ une fonction qui déplie une matrice en vecteur. Nous pouvons donc expliciter $\phi$ par :

$$\phi(B_i) = \text{vect}\left( \sum_r \mathbf{b}_{ri} \mathbf{b}_{ri}^\top \right). \quad (20)$$

Finalement, nous proposons de représenter un tube $i$ sous la forme d'un vecteur $\mathbf{t}_i = \frac{\phi(B_i)}{\|\phi(B_i)\|} \in \mathcal{H}$.

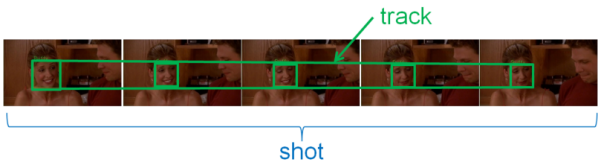


FIG. 1 – Exemple de tube vidéo décrivant un visage [11].

## 3.2 Signatures

**Signatures mono-tube.** Nous pouvons classifier les actions en utilisant directement la vidéo comme un tube vidéo. En effet, étant donné que chaque tube $i$ est décrit par un représentant $t_i$ dans un espace Hilbertien, on peut utiliser directement toutes les méthodes compatibles avec ces espaces, dans notre cas les SVM.

**Signatures multi-tubes.** Pour prendre en compte plus localement les mouvements, nous proposons d'extraire un ensemble de tubes vidéos (Figure 1). Nous proposons donc de définir pour chaque vidéo un ensemble de tubes vidéos. Chaque vidéo sera décrite par un sac de descripteurs de tube :

$$S_k = \{ \cdots, \mathbf{t}_{lk}, \cdots \} \in \mathcal{S} \quad (21)$$

avec $\mathbf{t}_{lk} \in \mathcal{H}$.

Nous présentons une approche pour extraire des tubes dans la partie 4.2.

Au même titre qu'avec les sacs $B$ de paires de frames, nous souhaitons utiliser des classifieurs par hyperplan. Nous proposons aussi d'effectuer un changement d'espace, cette fois-ci par le biais d'une fonction $\psi : \mathcal{S} \to \mathcal{C}$. Cependant, dans ce cas nous avons choisi d'utiliser une approche par codage [13]. Nous avons songé à réutiliser l'approche par

fonction noyau, mais la dimension des données crée des problèmes de complexité calculatoire.
Les méthodes d'indexation basées sur le codage sont inspirées des méthodes de compression utilisées en codage de source. En effet, ces méthodes cherchent à représenter les données avec un code plus court que la donnée d'origine tout en conservant le maximum d'information utile.
L'indexation par codage se déroulent en deux étapes :
– Une première étape, le *coding*, qui consiste en la transformation des descripteurs $\mathbf{t} \in \mathcal{H}$ en codes $c(\mathbf{t})$ dans un espace $\mathcal{C}$.
– Une deuxième étape, le *pooling*, qui consiste à rassembler tous les codes $\{c(\mathbf{t}_{lk})\}_r$ d'un même sac $S_k$ en un seul code $c(S_k)$.
J. Wang propose une méthode originale de coding, le Locality-Constrained Linear Coding (LLC) [13] qui utilise la contrainte suivante :

$$\begin{array}{l} \min\limits_{C} \sum\limits_{\mathbf{t}} \|\mathbf{t} - Dc(\mathbf{t})\|^2 + \lambda \|d(\mathbf{t}, D) \odot c(\mathbf{t})\|^2 \\ s.t.\ \forall \mathbf{t}\ \mathbf{1}^\top c(\mathbf{t}) = 1 \end{array} \tag{22}$$

où $D = (\mathbf{d}_j)_{j \in [1..M]}$ est un dictionnaire visuel, $C = \{c(\mathbf{t})\}_{\mathbf{t}}$ l'ensemble des codes, $\odot$ le produit de Hadamard et $d(\mathbf{t}, D)$ le vecteur de similarité entre le descripteur $\mathbf{t}$ et les vecteurs de la base $D$ :

$$d(\mathbf{t}, D) = \exp\left(\frac{\mathrm{dist}(\mathbf{t}, D)}{\sigma}\right) \tag{23}$$

avec $\mathrm{dist}(\mathbf{t}, D) = [\mathrm{dist}(\mathbf{t}, \mathbf{d}_1), \cdots, \mathrm{dist}(\mathbf{t}, \mathbf{d}_M)]^\top$ et $\mathrm{dist}(\mathbf{t}, \mathbf{d}_j)$ la distance euclidienne entre $\mathbf{t}$ et $\mathbf{d}_j$.
L'article [13] propose également une méthode approximée du LLC pour l'encodage rapide, cette méthode consiste à effectuer la projection d'un descripteur $\mathbf{t}$ sur un sous dictionnaire $D(\mathbf{t})$ spécifique à chaque descripteur $\mathbf{t}$. Ce sous-dictionnaire $D(\mathbf{t})$ est uniquement composé des $n$ plus proches descripteurs du dictionnaire $D$ de $\mathbf{t}$ (typiquement $n = 5$). On résout pour cela le problème suivant :

$$\begin{array}{l} \min\limits_{\tilde{C}} \sum\limits_{\mathbf{t}} \|\mathbf{t} - D(\mathbf{t})\tilde{c}(\mathbf{t})\|^2 \\ s.t.\ \forall \mathbf{t}\ \mathbf{1}^\top \tilde{c}(\mathbf{t}) = 1. \end{array} \tag{24}$$

Le dictionnaire de descripteurs $D$ est obtenu par l'application d'un algorithme de type K-Means sur l'ensemble des descripteurs de tube extraits de l'ensemble des vidéos. Par construction, chaque descripteur du dictionnaire est un barycentre d'un ensemble de descripteurs de tubes c'est pourquoi chaque descripteur du dictionnaire $\mathbf{d}_i$ peut être interprété comme le descripteur d'un tube type. Le nombre de barycentre du dictionnaire définit la taille du codage et donc la dimension du descripteur de sac que l'on obtient par cette méthode.
Il existe plusieurs méthodes pour effectuer l'opération de "pooling", par exemple :
– somme pooling : $c(S_k) = \sum_{\mathbf{t} \in S_k} \tilde{c}(\mathbf{t})$
– max pooling : $c(S_k) = \max_{\mathbf{t} \in S_k} \tilde{c}(\mathbf{t})$ .

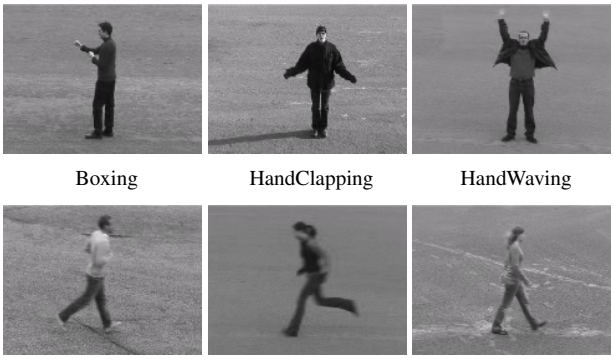


FIG. 2 – Exemple de vidéo de la base KTH

Les méthodes de "coding-pooling" permettent de définir un sac de descripteurs par un unique descripteur, tout en effectuant une forte compression des données et également de fournir une aussi grande importance à chaque types de descripteurs du sac quelque soit leur nombre d'occurrences.

# 4 Expériences

## 4.1 Base KTH

Pour tester le descripteur global nous utilisons la base de vidéos KTH [10]. Cette base contient six types d'actions humaines : walking, jogging, running, boxing, hand waving et hand clapping (Figure 2). Ces actions sont faites par 25 sujets différents dans quatre scénarios : dehors, dehors avec variation d'échelle, dehors avec différents vêtements, à l'intérieur.
Toutes les séquences sont en noir et blanc avec un fond homogène et une caméra statique de 25 images par seconde, elles ont une résolution de 160x120 et durent environ 4 secondes.

**Protocole expérimental.** Nous utilisons un classifieur SVM avec une fonction noyau triangulaire de $\sigma = 1$ avec une distance euclidienne. Le flot optique est estimé à l'aide de la méthode de Lucas-Kanade avec une fenêtre de taille 5.

**Résultats.** Le Tableau 1 présente les taux de reconnaissance pour plusieurs degrés de la base de polynômes.
Il est intéressant de remarquer que plus le degré de la base est grand, plus le descripteur est performant. Cela montre que même si le modèle de Legendre commence à être trop précis et à prendre en compte le bruit du flot optique, le classifieur SVM arrive à séparer le bruit des données.
Le meilleur résultat obtenu a été $79.04\%$ avec une base de degré 13. La matrice de confusion pour ce résultat est donnée dans le Tableau 2. Nous pouvons voir que les pires cas de confusion sont entre l'ensemble de mouvements running, jogging et walking. En fait, le mouvement jogging est un mouvement au milieu de walking et running : si une personne marche plus vite, son jogging sera plus rapide et il peut être confondu avec le mouvement running. De même,

| Degré de la base | Taux de reconnaissance |
|---|---|
| 1 | *0,638* |
| 2 | *0,675* |
| 3 | *0,709* |
| 4 | *0,737* |
| 5 | *0,754* |
| 9 | *0,787* |
| 13 | **0,790** |
| 15 | *0,776* |

TAB. 1 – Taux de reconnaissance pour différents degrés de la base de polynômes.

| | Box | HWav | HClap | Jog | Run | Walk |
|---|---|---|---|---|---|---|
| Box | 0,937 | 0,021 | 0,121 | 0,007 | 0,007 | 0,000 |
| HWav | 0,000 | 0,847 | 0,000 | 0,000 | 0,000 | 0,000 |
| HClap | 0,021 | 0,132 | 0,868 | 0,000 | 0,000 | 0,000 |
| Jog | 0,021 | 0,000 | 0,007 | 0,514 | 0,174 | 0,132 |
| Run | 0,021 | 0,000 | 0,000 | 0,201 | 0,722 | 0,014 |
| Walk | 0,000 | 0,000 | 0,000 | 0,278 | 0,097 | 0,854 |

TAB. 2 – Matrice de confusion avec une base de degré 13

si une personne marche plus lentement, son jogging sera plus lent et il peut être confondu avec le mouvement walking.
Si nous comparons les taux de reconnaissance avec différentes méthodes de la littérature (Tableau 3), nous pouvons voir que les taux sont du même ordre de grandeur.

## 4.2 Base Hollywood2

La base de test Hollywood2 [9] est constituée d'une collection de clips vidéo extraits de 69 films et répartis dans 12 classes d'actions humaines (Figure 3). Elle totalise approximativement 20 heures de vidéo et contient environ 150 échantillons vidéo par actions. Elle permet d'offrir une évaluation plus réaliste des méthodes de classification d'actions humaines en contenant des résolutions spatiales variées, des zooms de caméra, des scènes coupées et des artéfacts de compression.

**Protocole expérimental.** Nous utilisons un classifieur SVM avec un noyau gaussien et une validation croisée de $\sigma$ sur "test", la classification est de type "un contre tous" et les résultats sont la précision moyenne de classification.
Nous effectuons des tests de classification d'actions avec les deux méthodes présentées :

– **Mono-tube :** Pour la méthode mono-tube nous décrivons la vidéo par un unique descripteur du tube qui regroupe l'ensemble de la vidéo. Le flot optique a été calculé avec une taille de fenêtre de 75 pixels de coté et une base de polynôme de degré 6. Nous utilisons la distance euclidienne pour comparer les descripteurs de tube.
– **Multi-tubes :** Pour la méthode multi-tubes, nous décrivons la vidéo par un sac de descripteurs de tube, les tubes ont été sélectionnés par un découpage régulier de la vidéo avec recouvrement des tubes. Nous faisons varier les proportions du tube proportionnellement à celles de la vidéo, selon les plages présentées dans le tableau 4.

| Approche | Taux de reconnaissance |
|---|---|
| [5] | *0,940* |
| [1] | *0,912* |
| Notre approche | *0,790* |
| [10] | *0,717* |

TAB. 3 – Comparaison du taux de reconnaissance avec différentes méthodes de la littérature.


FIG. 3 – Exemple de vidéo de la base Hollywood2

| Dimension(s) | min | max | pas | Recouvrement |
|---|---|---|---|---|
| Spatial | 50% | 100% | 25% | 75% |
| Temporelle | 25% | 100% | 25% | 75% |

TAB. 4 – Plage de variation des dimensions des tubes vidéo.

Le flot optique a été calculé avec une taille de fenêtre de 20 pixels de coté et une base de polynôme de degré 3. On utilise un coding pooling avec une longueur de code de 4000 et la méthode de pooling : "max pooling". Nous utilisons la distance du $\mathcal{X}_2$ pour comparer les descripteurs de sac de tubes.

**Résultats.** Le tableau 5 représente la précision moyenne de classification pour chaque classe d'action.
Nous voyons que pour une classification des actions dans une base réelle la méthode mono-tube donne des taux de reconnaissance légèrement inférieurs à ceux trouvés dans la littérature mais elle est très simple et rapide à extraire et elle est également de petite dimension par rapport aux tailles des vidéos à classifier. La méthode mono-tube n'obtient pas de bon taux de classification dans les bases réelles car les actions à classifier sont "noyées" parmi des mouvements parasites, en effet la modélisation du flot optique par une base de polynôme va avoir tendance à principalement

| Méthode | CVPR09 [9] | Mono | Multi |
|---|---|---|---|
| AnswerPhone | *0.107* | *0.158* | **0.219** |
| DriveCar | **0.750** | *0.482* | *0.530* |
| Eat | **0.286** | *0.088* | *0.127* |
| FightPerson | *0.571* | *0.593* | **0.689** |
| GetOutCar | *0.116* | *0.111* | **0.206** |
| HandShake | *0.141* | *0.146* | **0.173** |
| HugPerson | *0.138* | *0.212* | **0.268** |
| Kiss | **0.556** | *0.371* | *0.398* |
| Run | **0.565** | *0.324* | *0.515* |
| SitDown | **0.278** | *0.243* | *0.249* |
| SitUp | **0.078** | *0.070* | *0.068* |
| StandUp | *0.325* | *0.282* | **0.454** |
| Moyenne | *0.326* | *0.256* | **0.328** |

TAB. 5 – Précision moyenne de reconnaissance des actions de la base Hollywood2.

prendre en compte les flots les plus forts qui ne sont pas forcément ceux qui caractérisent le mieux l'action.
Nous voyons que le résultat obtenu par la méthode multi-tubes est bien meilleur que les résultats de la méthode mono-tube. Elle permet en effet d'exploiter plus localement le flot optique dans la vidéo et donc d'éviter de "noyer" les actions locales. Cela a permis de dépasser le taux de reconnaissance moyen trouvé dans la littérature.
Nous pouvons remarquer que les actions qui créent de forts mouvements sur de longues périodes de temps comme les actions : DriveCar, FightPerson et Run, ont les meilleurs taux de reconnaissance.

# 5 Conclusion

Nous avons introduit une nouvelle méthode pour la reconnaissance d'actions dans les vidéos. Cette méthode est basée sur une technique qui modélise le flot optique entre deux frames successives de la vidéo à l'aide d'une base de polynômes orthogonaux. Puis nous avons présenté une méthode originale pour le calcul de signatures visuelles, ainsi qu'une métrique compatible avec la classification par SVM. Nous avons évalué cette méthode sur les bases vidéo de l'université de KTH et Hollywood2. Nous avons obtenu des résultats comparables avec ceux de la littérature. Nous travaillons actuellement sur l'automatisation efficace du réglage des paramètres de cette méthode, ainsi que sur la réduction de la taille des signatures.

# Références


[1] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, Cairo, Egypt, November 2009. IEEE Computer Society.

[2] Mert Dikmen, Huazhong Ning, Dennis J. Lin, Liangliang Cao, Vuong Le, Shen-Fu Tsai, Kai-Hsiang Lin, Zhen Li, Jianchao Yang, Thomas S. Huang, Fengjun Lv, Wei Xu, Ming Yang, Kai Yu, Guangyu Zhu, and Yihong Gong. Surveillance event detection. In *TRECVID*, 2008.

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

[4] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision*, pages 1395–1402, 2005.

[5] Nazli Ikizler, Ramazan Gokberk Cinbis, and Pinar Duygulu. Human action recognition with line and flow histograms. In *IAPR International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[6] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, jun 2008.

[7] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision (ijcai). In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI 81)*, pages 674–679, April 1981.

[8] S. Lyu. Mercer kernels for object recognition with local features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 223–229, San Diego, CA, 2005.

[9] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936, 2009.

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

[11] J. Sivic, M. Everingham, and A. Zisserman. Person spotting : Video shot retrieval for face sets. In *ACM International Conference on Image and Video Retrieval*, 2005.

[12] V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

[13] Jingjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3360–3367, 2010.