Reconnaissance d'actions dans des vidéos par caractérisation fréquentielle des trajectoires de points critiques

Cyrille Beaudry, Renaud Péteri, Laurent Mascarilla

Reconnaissance d'actions dans des vidéos par caractérisation fréquentielle des trajectoires de points critiques.

Cyrille Beaudry Renaud Péteri Laurent Mascarilla

Laboratoire Mathématiques, Image et Applications
Université de La Rochelle

{cyrille.beaudry, renaud.peteri, laurent.mascarilla}@univ-lr.fr

Résumé

Cet article porte sur la reconnaissance d'actions humaines dans des vidéos. La méthode présentée est basée sur l'estimation du flot optique dans chaque séquence afin d'en extraire des points critiques caractéristiques du mouvement. Des trajectoires d'intérêt multi-échelles sont ensuite générées à partir de ces points puis caractérisées fréquentiellement. Le descripteur final de la vidéo est obtenu en fusionnant ces caractéristiques de trajectoire avec des informations supplémentaires d'orientation de mouvements et de contours. Les résultats expérimentaux montrent que la méthode proposée permet d'atteindre sur la base KTH des taux de classification parmi les plus élevés de la littérature. Contrairement aux récentes stratégies nécessitant des grilles denses de points d'intérêt, la méthode a l'avantage de ne considérer que les points critiques du mouvement, ce qui permet une baisse du coût de calcul ainsi qu'une caractérisation plus qualitative de chaque séquence. Les perspectives de ce travail sont finalement discutées.

Mots Clef

Reconnaissance d'actions dans des vidéos, points critiques, caractérisation fréquentielle de trajectoires.

Abstract

This paper focuses on human action recognition in video sequences. A method based on the optical flow estimation is presented, where critical points of this flow field are extracted. Multi-scale trajectories are generated from those points and are frequentially characterized. Finally, a sequence is described by fusing this frequency information with motion orientation and shape information. Experiments show that this method performs on the KTH dataset this method achieves recognition rates among the highest in the state of the art. Contrary to recent dense sampling strategies, the proposed method only requires critical points of motion flow field, thus permitting a lower computational cost and a better sequence description. Results and perspectives are then discussed.

Keywords

Action recognition in videos, critical points, frequential characterization of motion trajectories.

1 Introduction

Dans le domaine de la détection de points d'intérêt spatio-temporels, Laptev et Lindberg [3] ont été les premiers à proposer une extension spatio-temporelle du détecteur de points d'intérêt 2D de Harris-Laplace. Dans [1], les auteurs proposent le détecteur *cuboïd*, basé sur des points d'intérêt calculés à partir de réponses de filtres de Gabor dans le domaine spatial et temporel. Une extension temporelle du détecteur de points d'intérêt 2D, basée sur le calcul de la Hessienne et permettant la détection de "blobs" dans les images est proposée dans [11]. Dans le cadre de la reconnaissance d'actions dans des vidéos, l'efficacité de la sélection dense de points à des positions régulières dans l'espace et dans le temps à différentes échelles a été démontrée [10]. On retrouve dans de nombreuses travaux le choix d'une sélection uniforme d'un très grand nombre de points d'intérêt plutôt qu'une estimation de points dits "rares". La contrepartie de ce choix est qu'il impose un coût de calcul beaucoup plus élevé en raison du grand nombre de points à traiter. Dans [9], les auteurs étendent cette selection dense en suivant la trajectoire de ces points sur un certain nombre d'images. [5] ou récemment [8] s'intéressent aussi à la caractérisation de trajectoire et à leur description afin d'en faire un élément discriminant performant. Ces récentes approches illustrent la pertinence du suivi de trajectoires de points pour la reconnaissance d'actions dans des vidéos.

Dans cet article, nous présentons une approche basée sur l'utilisation du flot optique et le suivi de points critiques de ce flot estimés à différentes échelles. L'idée est d'aller au delà du concept de point d'intérêt en considérant leurs trajectoires comme signature du mouvement. Ces trajectoires sont décrites en utilisant des coefficients de leur transformée de Fourier afin d'être invariantes en échelle, rotation, translation et robuste au bruit.

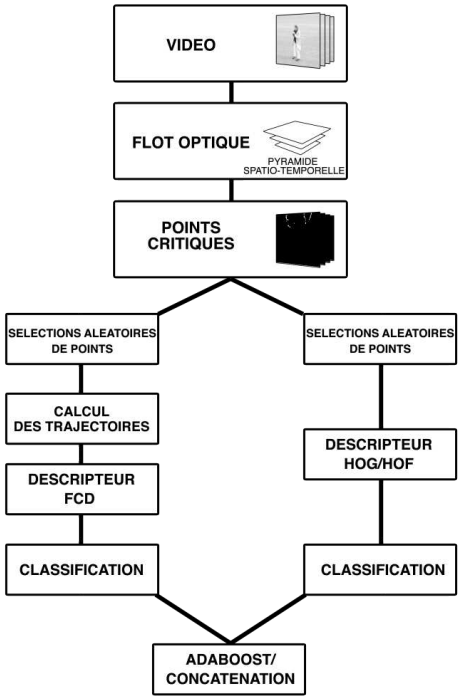

FIGURE 1 – Schéma général de l'approche proposée.

Cet article est organisé de la façon suivante. La section 2 détaille l'estimation des points critiques et des trajectoires multi-échelles. La section 3 présente la caractérisation des points critiques et des trajectoires multi-échelles, notamment l'utilisation des coefficients de transformée de Fourier pour ajouter une information fréquentielle à celles du mouvement et des contours. Enfin, en section 4 sont présentés des résultats expérimentaux sur la base KTH ainsi que des comparaisons avec l'état de l'art.

2 Points critiques et trajectoires

Cette section décrit la méthode permettant d'estimer les points critiques du flot optique et les trajectoires calculées à partir de ces points.

2.1 Points critiques d'un champ de vecteurs

L'estimation de points critiques est faite à partir d'une estimation robuste du flot optique utilisant un filtre médian à chaque itération [7]. Pour chaque image de la séquence, la divergence et le rotationel du flot optique sont calculés. Soit un champ de vecteur $\mathbf{F} = (u_t, v_t)$ avec u_t et v_t ses composantes horizontales et verticales, le rotationel et la divergence de F sont définis comme suit :

$$Rot(\mathbf{F}) = \nabla \wedge \mathbf{F} = \frac{\partial v_t}{\partial x} - \frac{\partial u_t}{\partial y}$$

$$Div(\mathbf{F}) = \nabla \cdot \mathbf{F} = \frac{\partial u_t}{\partial x} + \frac{\partial v_t}{\partial y}$$

Ces deux valeurs caractérisent la façon dont le champ de vecteurs évolue dans le temps :
- le rotationel donne une information sur la manière dont un champ de vecteur peut « tourner » localement.
- la divergence traduit le degré de comportement d'un point comme une source ou un puits pour le champ de vecteurs.
Les points avec une forte divergence ou un fort rotationel d'un champ de vecteurs sont caractéristiques d'une forte déformation locale. Nous utilisons ce critère pour trouver des zones de fortes déformations du flot optique, porteuses d'information sur de potentiels mouvements d'intérêt (Figure 2). Les points d'intérêt spatio-temporels sont donc les points extrema de rotationel et de divergence, qui correspondent à des points critiques du flot optique.

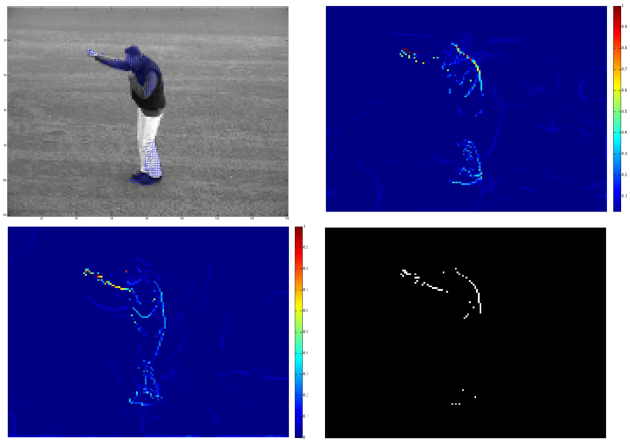

FIGURE 2 – Flot optique, points de courbure, points de divergence, extrema. Les points extraits correspondent localement et temporairement aux actions effectuées par le sujet dans la vidéo.

2.2 Extraction et caractérisation de trajectoires multi-échelles

Afin d'analyser les mouvements caractéristiques de la séquence, des trajectoires de mouvement sont calculées à partir des points critiques du flot optique obtenu en utilisant la méthode des trajectoires denses [9]. Ces points sont suivis tout le long de la séquence en appliquant un filtre median sur le flot optique. Etant donné le flot optique $F = (u_t, v_t)$, la position d'un point $P_t = (x_t, y_t)$ à l'image t, est estimée à l'image suivante $t+1$ par le point $P_{t+1} = (x_{t+1}, y_{t+1})$ selon :

$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + Med_F(V_{(x_t, y_t)})$

Avec Med_F, un filtre médian appliqué spatialement au flot F en $V_{(x_t, y_t)}$ qui est un voisinage du point P_t.
Afin d'analyser les différentes fréquences de mouvement pour les trajectoires extraites, une approche pyramidale spatio-temporelle a été retenue.
Une subdivision dyadique spatio-temporelle est effectuée sur les séquences. Les sous-séquences obtenues sont fil-

trées par un noyau gaussien spatio-temporel afin de supprimer les éventuelles hautes fréquences. Le flot optique est ensuite estimé sur chacune de ces séquences. Chaque sous-séquence correspond à une échelle de cette pyramide. La subdivision dyadique dans le temps permet d'obtenir des trajectoires de même longueur mais pour des fréquences temporelles différentes. Ce point est détaillé par la suite.
La déformation du flot peut être reliée à un mouvement possédant une échelle spatio-temporelle caractéristique. Les trajectoires issues de ce mouvement possèdent donc elles aussi une ou plusieurs fréquences caractéristiques. Dans cette idée, on obtient des mouvements correspondant à des trajectoires comprises dans un plus large intervalle de fréquences (Figure 3).

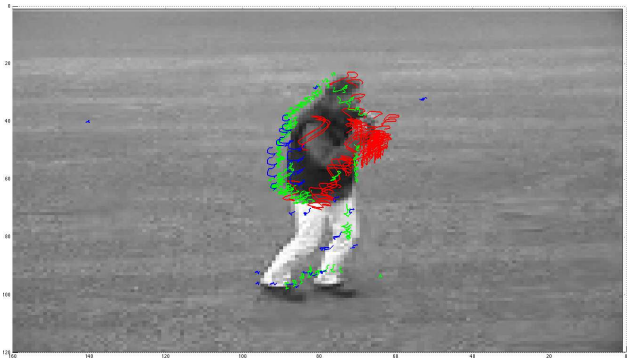

FIGURE 3 – Les trajectoires rouges sont extraites à partir des hautes fréquences du mouvement (poings), tandis que les vertes et les bleus correspondent à des mouvements de fréquences plus basses (jambes).

Nous suivons les points critiques extraits à chaque échelle de la pyramide afin de calculer des trajectoires d'échelles spatio-temporelles différentes, que nous appelons par la suite "trajectoires multi-échelles". La taille des trajectoires est proportionnelle à la durée des sous-séquences correspondantes de la pyramide. Pour une séquence de N images, la taille T_s des trajectoires est calculée comme suit :
$T_s = l_1 \cdot (2^{s-1}) \cdot N$
où s est l'échelle de l'étage de la pyramide et l_1 un seuil fixé empiriquement.
Quand la taille d'une trajectoire est plus grande que le seuil, elle est automatiquement tronquée afin de correspondre à la taille désirée. Si sa taille est plus petite que ce seuil, elle est supprimée. Cette condition permet de toujours garder de courtes trajectoires pour éviter les problèmes de non stationnarité durant le suivi, tout en permettant l'analyse. Toutes les trajectoires obtenues sont de même longueur mais correspondent à différentes fréquences de mouvement.

3 Descripteurs calculés à partir des points critiques et des trajectoires.

À partir des éléments calculés précédemment, c'est à dire les points critiques du flot optique et les trajectoires multi-

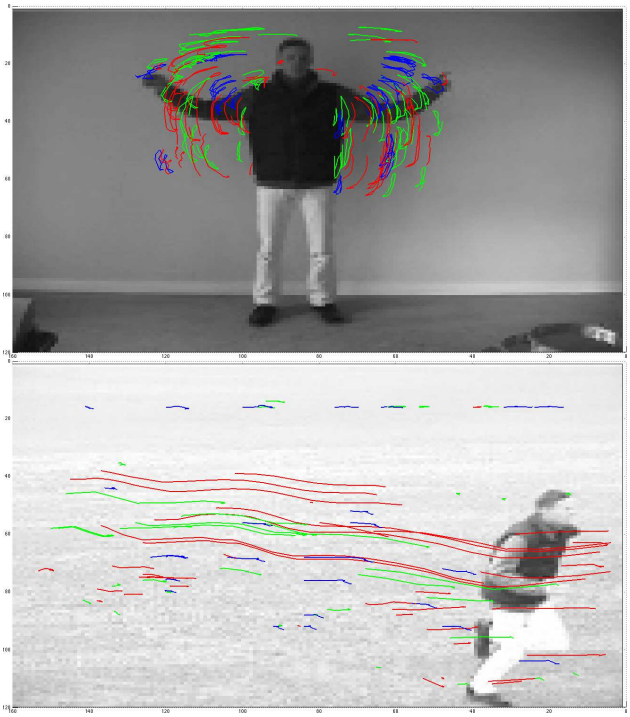

FIGURE 4 – Exemple de trajectoires extraites pour les actions "handwaving" et "running". On observe que dans les deux cas, les trajectoires correspondent aux mouvements présents dans la vidéo.

échelles estimées à partir de ces points, différents descripteurs sont extraits.

3.1 HOG/HOF

Le descripteur utilisé pour les points critiques est le descripteur classique HOG/HOF [10]. Il est à la fois basé sur l'information des contours (histogramme des orientation du gradient 2D) et l'orientation du mouvement (histogramme des orientations du flot optique). C'est un descripteur souvent employé dans la littérature car très performant.

3.2 Descripteur de trajectoires basé sur les coefficients de Fourier

Les trajectoires multi-échelles obtenues sont ensuite décrites par les coefficients de leur transformée de Fourier, ce qui permet d'utiliser l'information fréquentielle des mouvements comme élément discriminant de l'action dans la vidéo. Un système performant de reconnaissance d'action doit extraire des descripteurs possédant une faible variation intra-classe tout en assurant une invariance et une robustesse à différents types de transformations. Le choix des coefficients de Fourier est motivé par l'invariance à certaines transformations qu'il est possible d'obtenir dans le domaine fréquentiel (Figure 5).
Soit une trajectoire comportant N points séquentiels :
$T_N = [P_1, P_2, ..., P_t, ..., P_N]$
P_t étant un point quelconque de la trajectoire ayant comme

position (x_t, y_t, t).
Dans la suite, on considère que la transformée de Fourier d'une trajectoire T_N est $X_K = [X_0, X_1, ..., X_k, ..., X_{N-1}]$ tel que :
$X_k = \sum_{n=0}^{N-1} e^{\frac{-i2\pi kn}{N}} \cdot P_n$, $k \in [\![0, N-1]\!]$
avec le point $P_n = (x_n, y_n)$, et N la longueur de la trajectoire et k la fréquence d'analyse.

Pour obtenir l'invariance par translation, on soustrait aux coordonnées (x_n, y_n) des points de la trajectoire T_N leur valeur moyenne sur cette trajectoire :
$\tilde{x_n} = x_n - \sum_{t=1}^{N} \frac{x_t}{N}$ et $\tilde{y_n} = y_n - \sum_{t=1}^{N} \frac{y_t}{N}$
Afin d'obtenir une invariance par rotation, les trajectoires T_N sont traitées comme des vecteurs de nombres complexes et s'écrivent :
$T_{iN} = [P_{i1}, P_{i2}, ..., P_{it}, ..., P_{iN}]$
$P_{it} = \tilde{x}_t + i\tilde{y}_t$ étant la représentation complexe du point P_t.
Ainsi, pour une trajectoire $T_{\theta iN}$ représentant une rotation d'angle θ de la trajectoire initiale T_{iN}, la valeur absolue de la transformée de Fourier de $T_{\theta iN}$ et celle de T_{iN} sont égales. Il y a donc une invariance par rapport à la rotation.
L'invariance par rapport à l'échelle est assurée par la normalisation de la transformée de Fourier en divisant ses coefficients par la première composante fréquentielle non nulle. $\tilde{X_k} = \frac{X_k}{|X_0|}$, $k \in [\![0, N-1]\!]$
Finalement, le descripteur basé sur les coefficients de Fourier (FCD) est :
$FCD_{[T_{iN}]} = [\left|\tilde{X_0}\right|, \left|\tilde{X_1}\right|, ..., \left|\tilde{X_k}\right|, ..., \left|\tilde{X_{N-1}}\right|]$, $k \in [\![0; N-1]\!]$ tel que :
$X_k = \sum_{n=0}^{N-1} e^{\frac{-i2\pi kn}{N}} \cdot P_{in}$, $k \in [\![0, N-1]\!]$

Les trajectoires ayant toutes la longueur N, le descripteur FCD est donc calculé sur les mêmes plages fréquentielles $\frac{k}{N}$ avec $k \in [\![0, N-1]\!]$.
Les trajectoires sont ensuite lissées en supprimant les coefficients de transformée de Fourier correspondant aux très hautes fréquences, qui sont assimilées à du bruit ou des imprécisions de localisation. Ce traitement permet de rendre le descripteur robuste aux petites perturbations de mouvement (Figure 6).

4 Evaluation de la méthode pour la reconnaissance d'actions

Cette section décrit dans un premier temps la base de données utilisée pour la reconnaissance d'action. Par la suite, la méthode par sac de mots (bag of words) est présentée, ainsi que les paramètres utilisés pour notre approche.

4.1 Base de données utilisée

Base de données KTH. La base de données KTH [6] contient six classes d'actions humaines :"walking, jogging, running, boxing, waving, clapping". Chaque action est effectuée plusieurs fois par 25 sujets dans 4 scénarios différents. Toutes les séquences sont prises à 25 images/seconde avec un fond homogène et une caméra statique. La base de données contient en tout 600 videos.

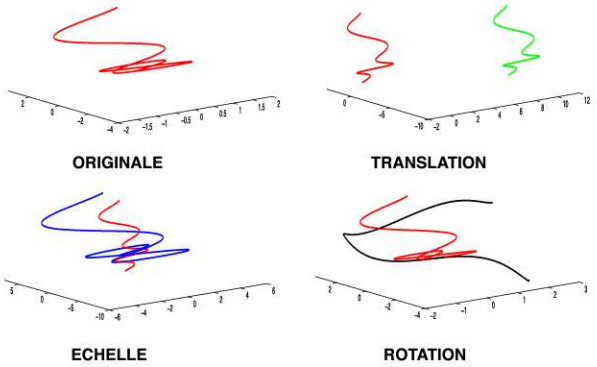

FIGURE 5 – Différentes transformations de la trajectoire originale (translation, échelle, rotation) donnant le même vecteur descripteur.

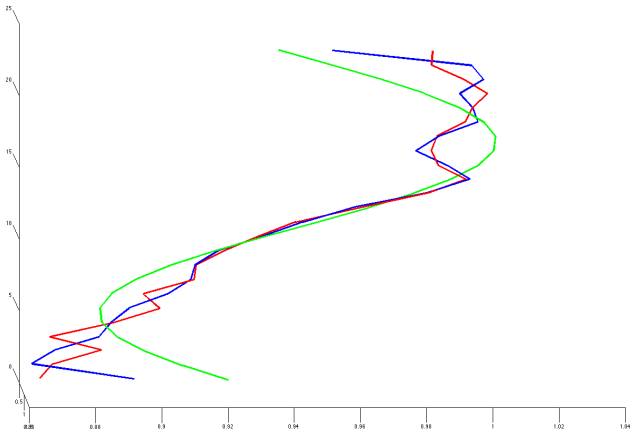

FIGURE 6 – Trajectoire originale (en rouge), trajectoire lissée en retenant 50% des coefficients (en bleu), trajectoire lissée en retenant 80% des coefficients (en vert).

4.2 Méthodologie

Approche par sac de mots visuels. Afin d'évaluer les performances de notre méthode pour la reconnaissance, l'approche dites de sac de mots [12] est utilisée. On considère que les vidéos de la base de données peuvent être décrites au moyen d'un dictionnaire de "mots visuels". La construction de ce dictionnaire se fait en partitionant, avec l'algorithme des k-moyennes, l'ensemble des vecteurs descripteurs calculés sur la base de données. Les centres obtenus forment les "mots visuels" du dictionnaire. Un vecteur descripteur est ensuite associé à son "mot visuel" le plus proche au sens de la distance euclidienne. Une vidéo est alors représentée par un histogramme d'occurrence de mots visuels du dictionnaire.

Cette méthode a montré son efficacité dans la reconnaissance de textes, d'images [12] et est désormais très utilisée dans la reconnaissance d'actions dans des vidéos.
L'approche dite "multi-canaux" [4, 9] est ensuite utilisée afin d'obtenir une version spatio-temporelle du sac de mots plus localisée. Elle consiste à subdiviser une vidéo en plusieurs cellules (spatiales et temporelles). Un histogramme de "mot visuel" est calculé sur chaque cellule. L'histogramme global de la vidéo est la concaténation des histogrammes de chacune de ses cellules. La subdivision de la vidéo en plusieurs cellules est appelée un canal. L'approche spatio-temporelle du sac de mots utilise différents canaux afin de combiner plus d'informations.
Dans la littérature [9], un canal c est noté $hx \times vy \times tz$ tel que :
-x correspond au nombre de subdivisions horizontales h.
-y correspond au nombre de subdivisons verticales v.
-z correspond au nombre de subdivisions temporelles t.

Étape de classification par SVM. Une classification supervisée de type SVM est mise en place avec un noyau gaussien multi-dimensionnel, permettant d'établir une distance entre des vidéos représentées par plusieurs histogrammes de différents canaux [12], et défini comme suit :
$K(x_i, x_j) = \exp(-\sum_{c \in C} \frac{1}{A_c} D(H_i^c, H_j^c))$
où H_i^c et H_j^c sont respectivement les histogrammes des vidéos x_i et x_j relatifs au canal c comme défini plus haut. $D(H_i^c, H_j^c)$ est la distance du χ^2 et A_c un coefficient de normalisation [12].
Un apprentissage est réalisé pour chaque descripteur. Deux méthodes sont employées pour la fusion des résultats. La première est la concaténation classique des différents canaux de chaque descripteur.
Une autre méthode est la fusion *a posteriori* avec la méthode Adaboost multiclasses [2].

4.3 Résultats

Nous évaluons ici les performances de notre méthode sur la base de données KTH, classiquement utilisée dans la reconnaissance d'actions.

Paramètres. Les calculs sont effectuées sur 120 images pour chaque vidéo. Le seuil l_1 fixant la taille des trajectoires est de 0.15, ce qui donne des trajectoires calculées sur 18 images.
L'influence du nombre de points critiques estimés et du nombre de coefficients de Fourier conservés pour le taux de reconnaissance moyen est évaluée sur un sous ensemble de la base (training). L'observation de la courbe ainsi obtenue (Figure 7) nous conduit à retenir un maximum de 500 points critiques et 80% de coefficients de Fourier.
L'approche multi-échelle apporte un gain en terme de résultats mais ce dernier n'est pas assez important par rapport au coût de calcul. Cela s'explique par le fait que la base KTH est construite dans un environnement contrôlé (caméra statique, fond homogène). De ce fait, l'utilisation d'une seule échelle de trajectoire s'est montré satisfaisante pour notre étude.
Pour l'approche multi-canaux du sac de mots, deux canaux sont utilisés : $h1 \times v1 \times t1$ et $h2 \times v1 \times t1$,

Comparaison entre descripteurs. Les taux de reconnaissance relatifs à notre approche sont présentés dans le Tableau 1. Le résultat final est obtenu en fusionnant les descripteurs FCD et HOG/HOF des deux façons présentées dans l'étape de classification.

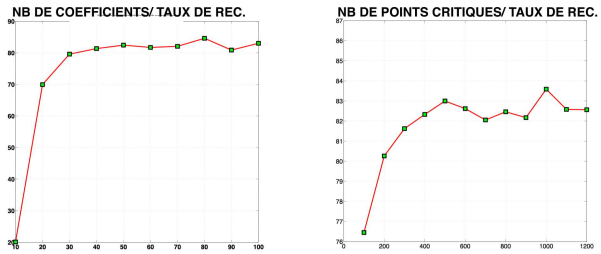

FIGURE 7 – Nombre de coefficients de Fourier gardés et nombre de points d'intérêt obtenus en fonction du taux de reconnaissance moyen. A partir d'un certain seuil, le taux de reconnaissance moyen évolue peu pour ces deux critères.

Descripteur	KTH Dataset
FCD	85.47%
HOG/HOF	91.98%
Combinaison	94.49%
Adaboost	**95.32%**

TABLE 1 – Taux de reconnaissance avec l'approche proposée pour différents types de descripteurs.

Discussion sur les résultats. Le descripteur FCD à lui seul donne un taux correct de reconnaissance mais est moins performant que le descripteur HOG/HOF seul. Cela peut s'expliquer par le fait que la base KTH ne peut être totalement discriminée fréquentiellement, et on constate que des actions tel que "boxing" et "handshaking" ont un contenu fréquentiel similaire (20.6% de confusion entre ces deux classes). Cependant, des actions tel que "jogging", "walking" et "running", proches visuellement mais s'effectuant à des fréquences différentes sont mieux discriminées par le descripteur FCD (4.6% de confusion entre ces trois classes). La méthode de fusion Adaboost (Figure 8) donne de meilleurs résultats que la concaténation des canaux des descripteurs : elle permet d'obtenir des taux de classification sur cette base supérieurs à la plupart des approches de l'état de l'art (Tableau 2). Cela illustre la complémentarité des informations combinées. Les approches récentes utilisant la notion de trajectoire [9, 5, 8], sont proches en terme de résultats. Elles sont néanmoins plus complexes à mettre

en oeuvre, notamment [5] qui utilise une méthode de suivi de motif pour le calcul de ses trajectoires ou encore [9] qui utilise une trentaine de canaux contrairement à notre approche qui n'en combine que deux.

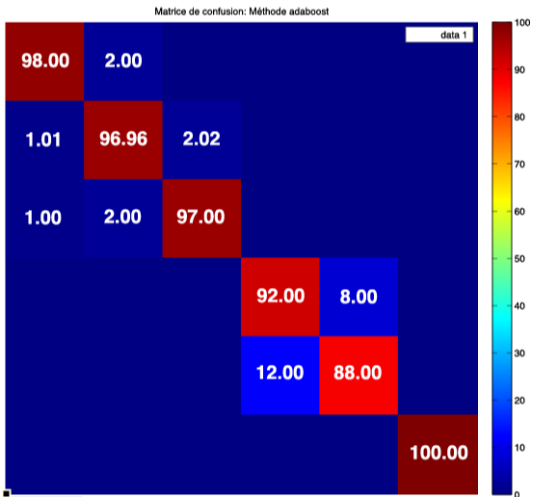

FIGURE 8 – Matrice de confusion obtenue après classification par SVM et fusion Adaboost

Méthode	**KTH Dataset**
Williems *et al.* [11]	88.7%
Dollar *et al.* [1]	89.1%
Laptev *et al.*[4]	92.1%
Wang *et al.* [9]	94.2%
Raptis *et al.* [5]	94.8%
Notre approche	**95.32%**
Vrigkas [8]	98.3%

TABLE 2 – Quelques taux de reconnaissance de la littérature sur la base KTH.

5 Conclusion

Ce papier présente une approche nouvelle de reconnaissance d'actions humaines dans des vidéos. Ces vidéos sont caractérisées par des points critiques calculés à partir du flot optique, ainsi que des trajectoires générées à partir de ce même flot.

Nos résultats montrent que l'information fréquentielle issue de ces trajectoires, combinée à l'orientation du mouvement et des contours présents au voisinage des points critiques permet d'obtenir des taux de reconnaissance parmi les plus élevés de la littérature sur la base KTH (Tableau 2), tout en utilisant une méthode de selection de points non dense, ce qui permet un temps de calcul réduit par rapport à d'autres approches similaires.

Les travaux en cours quant à l'application de notre méthode sur d'autres bases plus riches en actions et en informations montrent tout l'intérêt de l'approche multi-échelle en terme de gain quant à la combinaison des informations. Les résultats obtenus sont au niveau de l'état de l'art sur ces bases de données. L'amélioration de l'estimation des trajectoires ainsi que l'intégration d'informations contextuelles sont également des pistes en cours d'exploration.

Références

[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65 – 72, oct. 2005.

[2] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3) :349–360, 2009.

[3] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3) :107–123, September 2005.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR 2008*, pages 1 –8, june 2008.

[5] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *Proceedings of the 11th ECCV : Part I*, ECCV'10, pages 577–590, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions : a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36 Vol.3, 2004.

[7] D. Sun, S. Roth, and M.J. Black. Secrets of optical flow estimation and their principles. In *CVPR 2010. IEEE Conference on*, pages 2432–2439, 2010.

[8] M. Vrigkas, V. Karavasilis, C. Nikou, and A. Kakadiaris. Matching mixtures of curves for human action recognition. *CVIU*, 119(0) :27 – 40, 2014.

[9] H. Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR. 2011 IEEE Conference on*, pages 3169 –3176, june 2011.

[10] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *University of Central Florida, U.S.A*, 2009.

[11] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th ECCV : Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories : A comprehensive study. In *CVPR Workshop, 2006. CVPRW '06. Conference on*, pages 13–13, 2006.