Analyse multi-échelle de trajectoires de points critiques pour la reconnaissance d'actions humaines

Cyrille Beaudry, Renaud Péteri, Laurent Mascarilla

Laboratoire MIA, Univ. La Rochelle
Avenue Michel Crépeau
F-17042 La Rochelle CEDEX
{cyrille.beaudry,renaud.peteri,laurent.mascarilla}@univ-larochelle.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 différentes bases de vidéos, 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, notamment celle portant sur la reconnaissance d'actions dites complexes.*

ABSTRACT. *This paper focuses on human action recognition in video sequences. A method based on 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 on video datasets show that 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, comparison and perspectives on complex actions recognition are then discussed.*

MOTS-CLÉS : *reconnaissance d'actions, points critiques, caractérisation fréquentielle de trajectoires.*

KEYWORDS: *action recognition, critical points, frequential characterization of trajectories.*

DOI:10.3166/TS.32.265-286

1. Introduction

1.1. La reconnaissance d'actions humaines : une thématique active

La reconnaissance d'actions humaines dans des séquences d'images est une thématique de recherche dont l'intérêt est grandissant en vision par ordinateur, à la fois dans un contexte académique ou industriel. Le but est de pouvoir discriminer, dans des vidéos, différentes actions exécutées par un ou plusieurs sujets. Les algorithmes de reconnaissance d'actions sont préalablement entraînés avec une série d'exemples déjà étiquetés.

Beaucoup d'applications telles que l'indexation automatique de vidéos, la vidéo-surveillance, l'analyse de foule, l'interaction homme-machine ou l'analyse d'expressions faciales sont liées à une étape de reconnaissance d'actions. Ces problématiques très actuelles expliquent l'intérêt que suscite ce domaine de recherche. Ces différentes applications, de par leur contexte d'utilisation, amènent de nouvelles problématiques. Les séquences vidéos d'actions humaines contiennent généralement peu de contraintes d'acquisition, et présentent des changements de point de vue, des occultations partielles, des changements rapides d'illuminations, des mouvements de caméra, etc. Les données issues de ces vidéos, bien qu'elles puissent représenter des classes d'action communes sont donc extrêmement variables du fait de ces faibles contraintes. Le vocabulaire employé pour étiqueter ces vidéos est aussi d'une grande variabilité et dépend de la "granularité" sémantique de ce que l'on souhaite reconnaître. Ainsi, des actions telles que `ouvrir une porte` et `ouvrir un sac` sont souvent considérées comme deux classes différentes dans les bases de données, mais peuvent illustrer le même concept d'action dans le cas où l'objet de l'action effectuée importe peu. De plus, le contexte spatial où s'exécutent certaines actions peut parfois être l'élément le plus discriminant pour la reconnaissance (les actions `jouer du piano` et `jouer de la guitare` peuvent être complètement discriminées par l'instrument de musique). Une méthode efficace de reconnaissance d'actions se doit donc d'extraire toute information pertinente relative aux actions à reconnaître.

La démocratisation des appareils permettant la capture vidéo, tels que les smartphones, conduit à une augmentation constante de la taille des bases de données. 30 % du trafic internet est généré par des données videos. Les bases de données récentes prennent en compte ceci, en proposant un nombre croissant d'actions à discriminer ainsi qu'une masse de données de plus en plus importantes (plus de 2 millions d'images pour la base UCF-101).

Face à ces nouveaux challenges, les méthodes d'indexation automatique de vidéos doivent à la fois être efficientes et robustes mais aussi avoir des complexités calculatoires performantes. La reconnaissance d'actions humaines reste donc toujours une thématique parmi les plus dynamiques et complexes en vision par ordinateur.

1.2. Un bref état de l'art des méthodes de référence en reconnaissance d'actions

Dans la littérature, la reconnaissance d'actions humaines est abordée selon deux types de stratégies. Le premier type d'approche consiste en des méthodes génératives probabilistes, qui permettent une description de la structure temporelle des actions présentent dans des vidéos. La plupart des méthodes génératives probabilistes pour la reconnaissance d'activités s'appuient sur l'allocation de Dirichlet latente (LDA). Ce modèle permet de trouver les thèmes sous-jacents d'un document, et dans un cadre applicatif plus général, de caractériser une donnée (texte, image, séquence vidéo, etc) comme la proportion de thèmes (appelés topics) dont elle est composée. Cependant, les approches basées sur des modèles génératifs sont majoritairement non supervisées. Les thèmes sous-jacents d'une vidéo sont découverts automatiquement par l'algorithme LDA. Dans le cadre de la reconnaissance d'actions humaines, les actions exécutées dans les bases de données sont déjà connues, chaque vidéo est donc préalablement étiquetée. Ceci rend les modèles génératifs probabilistes globalement peu efficaces pour ce qui est du taux reconnaissance dans la reconnaissance d'actions humaines.

La seconde approche repose sur des méthodes basées sur des modèles discriminatifs. Ces méthodes sont très utilisées dans ce cadre parce qu'elles ont montré leur efficacité dans des domaines connexes, notamment dans la recherche de documents et la reconnaissance d'objets dans des images. Les méthodes proposées dans ce domaine suivent une méthodologie récurrente. En premier lieu, elles comprennent une phase de sélection de points d'intérêts susceptibles d'être discriminants pour le ou les actions présentes dans la vidéo. Ensuite, des descripteurs issus de ces points d'intérêt, sont calculés pour caractériser la vidéo traitée. Enfin, une phase de classification, entraînée sur une base d'apprentissage, et utilisant les vecteurs de descripteurs obtenus, permet la reconnaissance d'une ou des actions présentes dans la vidéo. Les principales différences entre les méthodes se situent donc au niveau de l'extraction des points d'intérêts et de leurs descripteurs, ainsi que dans les stratégies d'utilisation de ces descripteurs lors de la phase de classification.

La grande majorité des premières méthodes d'extraction de points d'intérêt spatio-temporels sont issues d'extensions dans le temps de détecteur de points d'intérêt 2D. (Laptev, 2005) ont été les premiers à extraire des points d'intérêt spatio-temporels (STIP) à partir d'une extension temporelle du détecteur de points 2D Harris-Laplace. Dans (Dollar *et al*., 2005), les auteurs proposent le détecteur *cuboïd*, basé sur des points d'intérêt calculés à partir de réponses à des filtres de Gabor dans le domaine spatial et temporel. Une extension temporelle du détecteur de points d'intérêt dans des images, basée sur le calcul de la Hessienne et permettant la détection de "blobs" est proposée dans (Willems *et al*., 2008). Chacune de ces méthodes passe par la maximisation d'une fonction caractérisant la présence d'évènements temporels donnés (blobs, coins, fréquences, etc). D'autres auteurs montrent par la suite que dans le cas de vidéos plus réalistes et génériques (extraits de films, vidéos issues de Youtube), les points

d'intérêt des détecteurs de la littérature prennent mal en compte la complexité et la richesse des informations de ces vidéos. Ainsi, l'efficacité de la sélection dense de points d'intérêt à des positions régulières dans l'espace et le temps et pour différentes échelles a été démontrée dans (Wang *et al*., 2009). Le choix d'une sélection uniforme d'un très grand nombre de points plutôt que d'une estimation de points dits d'intérêt a été par la suite repris. Dans (Wang *et al*., 2011), les auteurs étendent cette sélection dense et suivent la trajectoire de ces points sur plusieurs images consécutives. Le concept de trajectoires d'actions encode l'information temporelle de façon plus naturelle que le simple point d'intérêt. (Raptis, Soatto, 2010) ou très récemment (Vrigkas *et al*., 2014) mettent en valeur la pertinence du suivi de trajectoires de points d'intérêt pour la reconnaissance d'actions dans des vidéos. Dans (Laptev *et al*., 2008), les auteurs renoncent à l'étape de sélection automatique d'échelle pour le détecteur STIP, issue de (Laptev, 2005), en privilégiant une sélection dense en espace et en temps des points d'intérêt. Cette sélection dense des points d'intérêt permet d'accroître les performances de la méthode pour la reconnaissance d'actions dans des vidéos réalistes. Ces travaux seront ensuite repris dans (Ullah, Laptev, 2012) avec l'ajout de la caractérisation des trajectoires d'actions. A l'aide de bases de données de vidéos synthétiques, les mouvements de parties du corps humain exécutant certaines actions sont appris, puis ces mouvements sont retrouvés dans des vidéos plus génériques n'ayant pas servi dans la phase d'apprentissage.

Les méthodes denses font actuellement partie des approches les plus performantes, mais souffrent toutefois du même inconvénient : elles conduisent à des temps de calculs et un nombre de descripteurs très importants. La taille des données générées et le temps de calcul qu'impliquent les grandes bases de données actuelles réduisent les possibilités d'implémentation temps-réel basées sur ces méthodes denses.

Le critère du nombre de données à générer pour obtenir des résultats satisfaisants fait partie des travaux de récentes études. (Shi *et al*., 2013) montrent que l'utilisation d'un nombre fixe de caractéristiques sélectionnées dans un ensemble dense permet d'obtenir des résultats proches de l'état de l'art sur certaines bases de données génériques. À l'aide d'une méthode d'extraction de caractéristiques par grille dense, les caractéristiques sont aléatoirement choisies toutes les 160 trames pour en conserver un total de 10 k. Cela permet de conserver un grand nombre de points sur les échelles les plus fines tout en contrôlant leur nombre. Les résultats obtenus sur des grandes bases de données comme HMDB51 donne un gain de 1 % par rapport aux méthodes denses de l'état de l'art, mais restent peu élevés sur d'autres bases de données grande taille telles que UCF-50. (Murthy, Goecke, 2013) proposent une méthode de sélection des trajectoires obtenues par la méthode des trajectoires denses (Wang *et al*., 2011). Cette approche permet, avec environ deux fois moins de trajectoires et à paramètres équivalents, d'obtenir des taux de reconnaissances élevés sur des bases de données comme UCF-50. Cependant cette étape de sélection de trajectoires se fait après la génération de toutes les trajectoires denses. Elle ne fait donc pas l'économie du temps de calcul et du stockage temporaire de ces dernières.

1.3. Contribution de l'article

L'approche présentée tente de répondre aux différentes problématiques actuelles de la reconnaissance d'actions humaines dans les vidéos : performance face à différents types d'actions, robustesse face à la variabilité des données et de l'acquisition (mouvements de caméra, rotation, translation, etc.), utilisation de caractéristiques pertinentes, génération d'un faible taux de caractéristiques pour optimiser les temps de calculs. Nous faisons ici le choix de signer les actions dans les vidéos par le mouvement. L'estimation du flot optique, en tant que méthode courante pour estimer le mouvement entre deux images consécutives, est la base de cette approche. Cette estimation se fait en prenant en compte les mouvements parasites de la caméra, présents dans la grande majorité des vidéos sans contraintes d'acquisitions. Cette compensation du mouvement de la caméra se fait lors de l'estimation du flot optique, permettant ainsi de ne pas entraîner de coût additionnel en temps de calcul. Les points critiques du flot optique ainsi que leur trajectoire, estimées au cours du temps pour différentes échelles, sont les éléments d'intérêts de la méthode. Ces trajectoires sont caractérisées fréquentiellement pour une description robuste des mouvements au cours du temps. A cette information de fréquence s'ajoutent des descripteurs d'orientation de mouvement et de forme, très caractéristiques des mouvements humains. Cette approche permet d'obtenir des taux de reconnaissances sur des bases de données de vidéos avec contraintes d'acquisition, et de vidéos génériques, parmi les plus élevés de la littérature, tout en permettant de garder une complexité plus faible que les stratégies denses récentes, en temps et en espace de stockage.

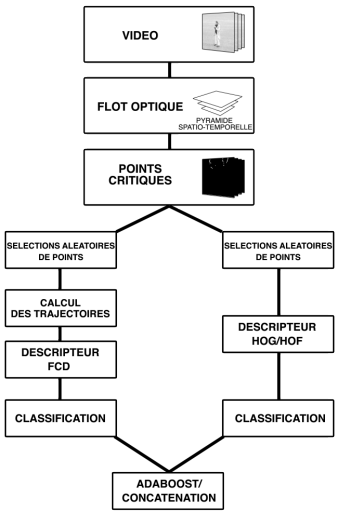

Figure 1. Schéma général de l'approche proposée

Cet article est organisé de la façon suivante. La section 2 détaille l'estimation des points critiques, des trajectoires multi-échelles et de la compensation du mouvement

de caméra. 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 d'orientation du mouvement et des contours. Enfin, en section 4 sont présentés des résultats expérimentaux sur différentes bases de données de la littérature ainsi que des comparaisons avec l'état de l'art.

2. Points critiques et trajectoires

2.1. Points critiques d'un champ de vecteurs

Nous estimerons le flot optique pour caractériser par le mouvement les actions présentes dans des séquences vidéos. Pour chaque image de la séquence, la divergence et le rotationel du flot optique sont calculés.
Soit un champ de vecteur $\mathbf{F_t} = (u_t, v_t)$ avec u_t et v_t ses composantes horizontales et verticales, le rotationel et la divergence de $\mathbf{F}$ sont définis comme suit :

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

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

Ces deux fonctions 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 mesure à quel degré un point du champ de vecteur est une source ou un puits.

Les extrema de ces deux composantes correspondent à certains points critiques du flot optique (tourbillons, points d'attraction, points de repulsion). Ces points sont caractéristiques de fortes déformations locales du champ vectoriel de la séquence et sont donc porteurs d'informations sur de potentiels mouvements d'intérêt (Figure 2).

2.2. Extraction et caractérisation de trajectoires multi-échelles

2.2.1. Trajectoires des points critiques

Pour aller au-delà de la simple notion de points d'intérêts, des trajectoires de mouvement sont calculées, à partir des points critiques du flot optique, par la méthode des trajectoires denses (Wang *et al.*, 2011). Ces points sont suivis tout le long de la séquence. Les points critiques estimés étant le point de départ de ces trajectoires, ces dernières sont caractéristiques des mouvements humains potentiels de la séquence. Étant

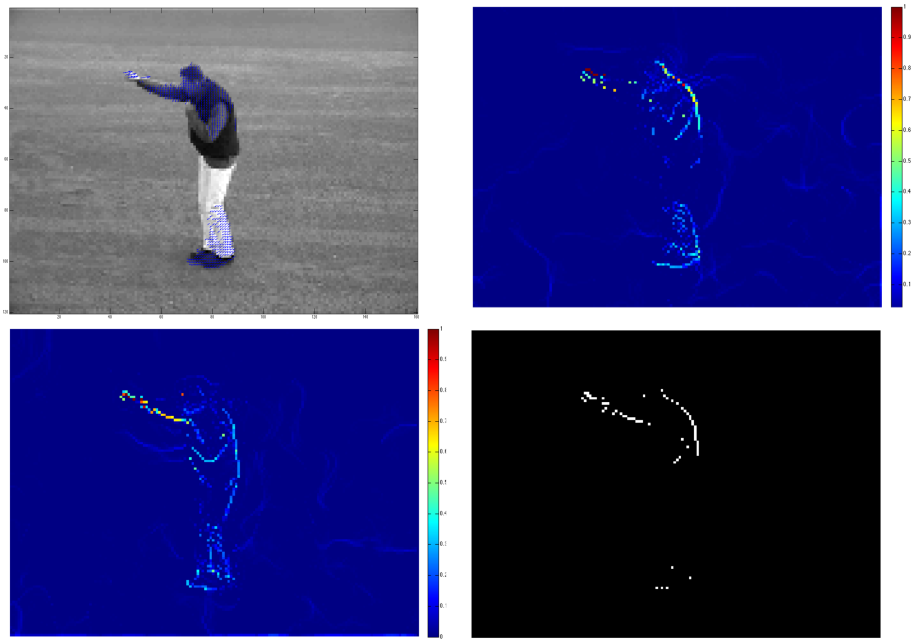

Figure 2. De haut en bas et de gauche à droite : Flot optique, points de rotation, points de divergence, extrema. Les points extraits correspondent localement et temporellement aux actions effectuées par le sujet dans la vidéo

donné le flot optique $\mathbf{F}_t = (u_t, v_t)$, la position d'un point $P_t = (x_t, y_t)$ à l'image t, est estimée à l'image suivante $t+1$ comme étant le point $P_{t+1} = (x_{t+1}, y_{t+1})$ tel que:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + Med_{\mathbf{F}_t}(V_{(x_t,y_t)})$$

avec Med_F, le filtre médian spatial appliqué sur le flot $\mathbf{F_t}$ au voisinage $V_{(x_t,y_t)}$ de P_t.

2.2.2. Caractérisation multi-échelle des trajectoires

Afin d'analyser les différentes fréquences de mouvement pour les trajectoires extraites, une approche pyramidale a été retenue. Une subdivision dyadique est effectuée sur chaque séquence en espace et en temps. Les sous-séquences obtenues sont filtrées par un noyau gaussien spatio-temporel afin de supprimer les 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 spatiale permet d'estimer des points critiques à différentes échelles spatiales. 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.

Nous suivons les points critiques extraits à chaque niveau de la pyramide afin de calculer des trajectoires d'échelles spatio-temporelles différentes, que nous appelons par la suite "trajectoires multi-échelles". Quand la taille d'une trajectoire est plus

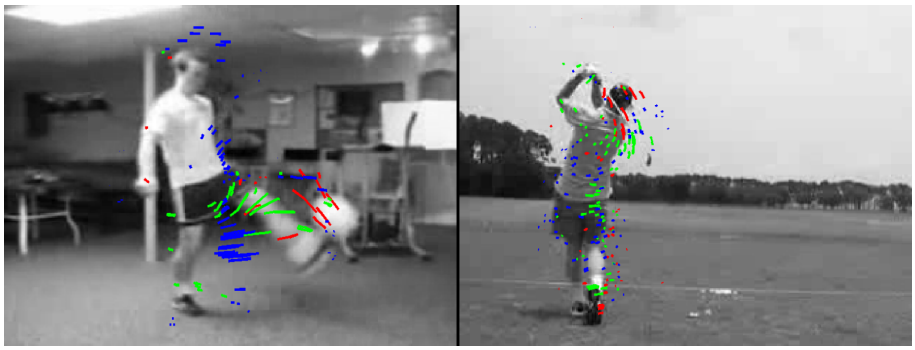

Figure 3. Exemple de trajectoires multi-échelles. Les trajectoires rouges correspondent à l'échelle dyadique la plus haute (mouvements cours et rapides). Les trajectoires bleus à l'échelle la plus basse (mouvements de basses fréquences). Les trajectoires vertes correspondent à une échelle dyadique intermédiaire. Sur la première figure on observe les trajectoires rouges au niveau des pieds du joueur et de la balle. Les trajectoires vertes correspondent au mouvement de la jambe et les bleues correspondent au mouvement global du corps

grande qu'un seuil fixé, elle est 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 des trajectoires courtes et d'éviter les problèmes de non-stationnarité durant le suivi, tout en permettant une analyse fiable. Les tests montrent que des trajectoires estimées sur une durée de 16 trames permettent d'obtenir de bons résultats de reconnaissance sur la plupart des bases de données. Toutes les trajectoires obtenues sont de même longueur mais correspondent à différentes fréquences de mouvement.

La déformation locale du flot optique peut être liée à un mouvement possédant une échelle spatio-temporelle caractéristique. Les trajectoires issues de ce mouvement sont donc elles aussi associées à une ou plusieurs fréquences caractéristiques. De ce constat, on obtient des trajectoires correspondant à des mouvements compris dans un plus large intervalle de fréquences (Figure 3). Cette approche nous permet d'isoler les trajectoires associées à différentes fréquences de mouvement. Les premières échelles dyadiques révèlent les mouvements les plus rapides (trajectoires courtes et hautes fréquences), tandis que les échelles les plus basses mettent en évidence les mouvements les plus lents (trajectoires longues et basses fréquences). Ceci permet d'obtenir une bonne caractérisation des trajectoires et des mouvements présents dans la séquence vidéo.

2.3. Compensation des mouvements de caméra

La difficulté dans l'estimation des trajectoires de points critiques est de garder une faible erreur d'estimation de leur position au cours du temps. Dans le cas de vidéos non contraintes, cette estimation peut être faussée du fait de nombreux mouvements

de caméra ou de changements de point de vue. Ces mouvements de caméra influent directement sur la qualité de l'estimation et donc sur la pertinence des trajectoires.

La disponibilité de bases de données récentes comprenant des vidéos en conditions non contraintes, fait que la correction du mouvement de la caméra est une étape de plus en plus importante. Parmi les différentes stratégies de la littérature pour traiter cette problématique, (Wang, Schmid, 2013) modélisent les mouvements de caméra entre deux images consécutives de la sequences par une homographie. L'estimation des paramètres de cette homographie se fait par correspondance de points d'intérêts locaux de type SURF, robustes au flou de bougé. Cette démarche permet d'obtenir un gain de 2,6 % sur UCF-50 avec un taux de reconnaissance de 91,2 % (tableau 3). En contrepartie, cette approche rajoute une complexité non négligeable de par l'utilisation d'un processus *ad hoc* de detection automatique d'humains dans des images et l'utilisation de la méthode RANSAC pour l'estimation des paramètres de l'homographie.

(Jain *et al.*, 2013) supposent que le mouvement d'une séquence vidéo peut être séparé en deux composantes : le mouvement dominant, dû à la caméra, et le mouvement résiduel, relatif aux actions présentes dans la séquence. Le mouvement dominant est extrait par estimation du flot affine entre deux images consécutives. La compensation de caméra est alors obtenue en soustrayant le flot affine au flot optique. Cette approche présente de bons résultats sur les bases de données récentes d'actions humaines, mais entraîne l'estimation de deux champs vectoriels, le flot optique ainsi que le flot affine, censé contenir l'information de mouvement de la caméra. La méthode que nous proposons utilise cette supposition sans estimer le mouvement dominant par un champ vectoriel supplémentaire.

2.3.1. Estimation du mouvement global par approche pyramidale

Afin de minimiser l'effet du mouvement de caméra tout en gardant un temps de calcul faible et en évitant des méthodes *ad hoc* trop coûteuses, nous tirerons parti du flot optique déjà estimé précédemment (section 2.2.1). Plus précisément, nous utiliserons une estimation multi-échelle du flot optique pour compenser le mouvement global de caméra. Ainsi, le champ de déplacement à l'instant t entre deux échelles I^l et I^{l+1} de la pyramide est tel que :

$$\mathbf{F_t^l} = E_2(\mathbf{F_t^{l+1}}) + f([I_t^l + E_2(\mathbf{F_t^{l+1}})], I_{t+1}^l)$$

avec l l'échelle dans la pyramide (l'échelle maximale est fixée à $L = 4$ par la suite), E_2 un opérateur d'interpolation qui permet le passage d'une matrice de taille $k \times k$ en une matrice de taille $2k \times 2k$, et f le flot optique estimé entre deux images. À l'échelle maximale L, le champ vectoriel estimé correspond aux mouvements les plus larges dans l'image, souvent dus aux mouvements de la caméra. Similairement à (Jain *et al.*, 2013), le mouvement de caméra est compensé directement durant le processus d'estimation du flot optique. Nous obtenons donc :

$$\mathbf{F^0}_{comp} = \mathbf{F^0}_{original} - \mathbf{F^L}_{original}$$

avec $\mathbf{F^0}_{original}$ l'estimation du flot optique entre deux images consécutives et $\mathbf{F^L}_{original}$ l'estimation du flot optique obtenu à la dernière échelle de la représentation pyramidale des images. Il modélise le mouvement global de la caméra. $\mathbf{F^0}_{comp}$ représente donc le flot optique entre les deux images consécutives avec la compensation du mouvement de la caméra.

La Figure 4 illustre le résultat obtenu sur une vidéo séquence issue de la base UCF-11 (Liu *et al.*, 2009). Nous verrons dans la partie expérimentale que cette approche améliore sensiblement la qualité des descripteurs de trajectoires (voir tableau 2).

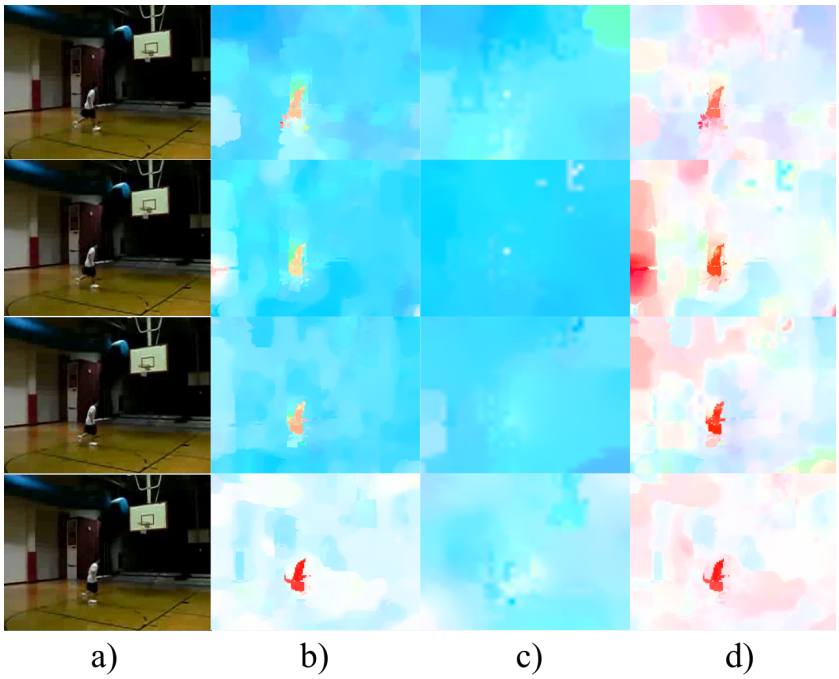

Figure 4. a) : quatre images consécutives avec un mouvement de caméra latéral sur les trois premières images. b) : estimation du flot optique $F^0_{original}$. c) : estimation du mouvement global $F^N_{original}$, d) : Compensation du mouvement de caméra. F^0_{comp}

L'estimation du mouvement global de la caméra se faisant directement lors de l'estimation du flot optique, notre approche permet une compensation sans coût de calculs supplémentaires.

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

Les points critiques obtenus ainsi que leurs trajectoires multi-échelles sont ensuite décrits par trois informations caractéristiques du mouvement : l'orientation, la varia-

tion de la forme autour des points critiques ainsi que l'information fréquentielle de ce mouvement.

3.1. Descripteur de trajectoires basé sur les coefficients de transformée de Fourier

3.1.1. Analyse fréquentielle des trajectoires

Les trajectoires multi-échelles obtenues sont ensuite décrites par les coefficients de leur transformée de Fourier, afin d'utiliser l'information fréquentielle des mouvements comme élément discriminant de l'action dans la vidéo. Le choix des coefficients de Fourier est de plus motivé par l'invariance à certaines transformations que l'on peut obtenir dans le domaine fréquentiel (Figure 5). Un système performant de reconnaissance d'actions doit en effet extraire des descripteurs possédant une faible variation intra-classe tout en assurant une invariance et une robustesse à différents types de transformations.

3.1.2. Invariances du descripteur

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) à l'instant t.

Dans la suite, on considère que la transformée de Fourier d'une trajectoire T_N est $X_K = [X_0, X_1, ..., X_k, ..., X_{N-1}]$ telle que :
$X_k = \sum_{t=0}^{N-1} e^{\frac{-i2\pi kt}{N}} \cdot P_{t+1}$, $k \in [\![0, N-1]\!]$

avec le point $P_t = (x_t, y_t)$, 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 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 :

$X_k = \sum_{t=0}^{N-1} e^{\frac{-i2\pi kt}{N}} \cdot P_{i(t+1)}$, $k \in [\![0, N-1]\!]$

Les trajectoires ayant toutes la longueur N, le descripteur FCD est calculé sur les mêmes plages fréquentielles $\frac{k}{N}$ avec $k \in [\![0, N-1]\!]$.

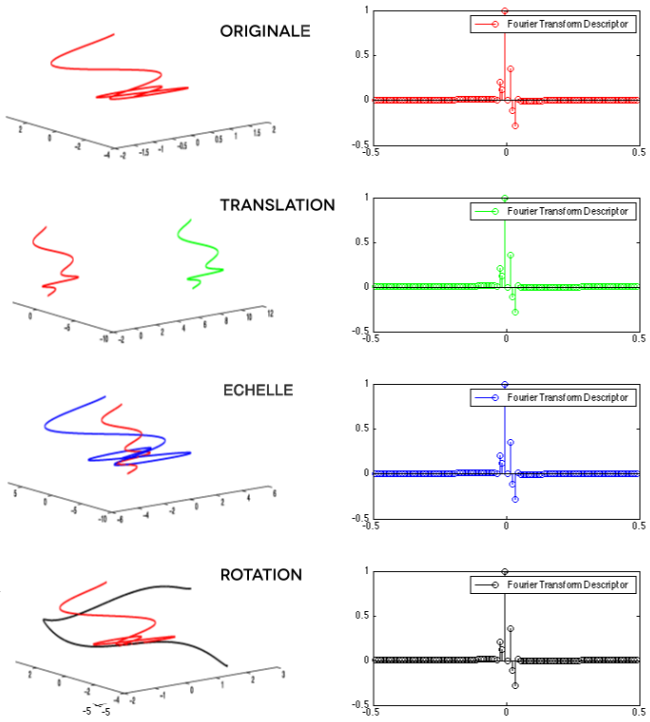

Figure 5. Différentes transformations géométriques de la trajectoire originale (translation, échelle, rotation) aboutissant au même vecteur descripteur

Les trajectoires sont ensuite lissées en supprimant les coefficients de la 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).

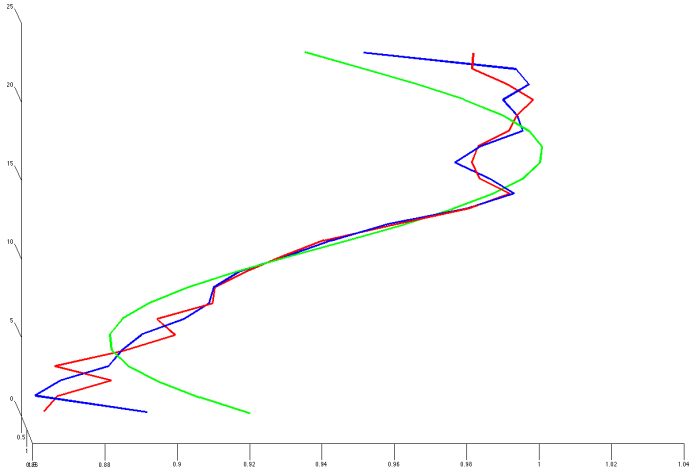

Figure 6. Trajectoire originale (en rouge), trajectoire lissée en supprimant 50 % des coefficients (en bleu), trajectoire lissée en supprimant 80 % des coefficients (en vert)

3.2. Caractérisation des variations de formes et d'orientation du mouvement.

Les descripteurs HOG (Histogram Of 2D Gradient) et HOF (Histogram of Orientation of optical Flow) sont utilisés pour caractériser les points critiques et leur trajectoires (Laptev *et al.*, 2008).

Le descripteur HOG décrit l'évolution du gradient 2D dans le voisinage d'un point critique. Il permet d'encoder l'information de contour et de forme des mouvements présents dans la séquence.

Le descripteur HOF caractérise l'information d'orientation du flot optique autour d'un point critique. Ce descripteur est souvent employé dans la littérature car très performant. Il décrit la variation locale de l'orientation du flot optique, qui est une information très caractéristique du mouvement.

La variation des contours, l'orientation du mouvement ainsi que l'information fréquentielle, sont des informations peu corrélées et donc complémentaires. En effet, le descripteur HOG est basé sur les variations locales du gradient, le descripteur HOF est lié à l'estimation du flot otique et le descripteur FCD caractérise les différentes fréquences de mouvement présentes au cours du temps.

Nos expérimentations, exposées par la suite, montrent à la fois la pertinence de ces informations dans le processus de classification, tout en exploitant la non-corrélation de ces caractéristiques par une fusion tardive de l'information.

4. Évaluation de la méthode pour la reconnaissance d'actions

Cette section décrit dans un premier temps les bases de données utilisées pour la reconnaissance d'actions. Par la suite, la méthode par sac de mots visuels (*bag of visual words*) est présentée, ainsi que les paramètres utilisés sur les différentes bases de données.

4.1. Bases de données utilisées

Nous utilisons ici à la fois des bases de données fortement contraintes au niveau de l'acquisition (caméra statique, fond quasiment homogène, etc.) mais aussi des bases plus génériques, avec de faibles contraintes d'acquisition, issues notamment de Youtube (changements de point de vue, occultations partielles, séquences couleurs, caméra non fixe,...). Les bases de données en conditions contrôlées permettent dans un premier temps d'évaluer l'efficacité de la méthode et de se comparer à l'état de l'art. Les bases de données plus récentes en conditions non contrôlées permettent de tester la robustesse de la méthode.

4.1.1. KTH Dataset

La base de données KTH (Schuldt *et al.*, 2004) 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 vidéos.

4.1.2. Weizmann Dataset

La base de données Weizmann (Gorelick *et al.*, 2007) est une collection de 90 vidéos avec les mêmes contraintes d'acquisitions et sans mouvement de caméra. Elle comptabilise un total de 10 actions différentes avec plusieurs actions similaires telles que : *jack, run, skip, side*.

4.1.3. UCF-11 Dataset

La base UCF-11 (Liu *et al.*, 2009) est une base de données construites à partir de vidéos provenant de Youtube. C'est une base de vidéos avec peu de contraintes d'acquisition (mouvements de caméra, zooms, changements de point de vue, occultations partielles, changements d'illumination, etc). Elle comptabilise un total de 11 actions : *basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, walking with a dog*.

4.1.4. UCF-50 Dataset

UCF-50 (Reddy, Shah, 2012) est une extension de UCF-11 avec plus de 50 catégories d'actions contenues dans des vidéos provenant de Youtube. Les difficultés de cette base de données résident à la fois dans la grande variabilité des informations visuelles liées aux différentes actions, le nombres d'actions à classer ainsi que le nombre de vidéos à traiter (5 000 videos).

4.2. Méthodologie de l'approche proposée

4.2.1. Approche par sac de mots visuels

Afin d'évaluer les performances de notre méthode pour la reconnaissance, l'approche dite des "sacs de mots visuels" (Lazebnik *et al.*, 2006) 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 partitionnant, généralement 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 (Lazebnik *et al.*, 2006) et est désormais couramment utilisée dans la reconnaissance d'actions dans des vidéos. L'approche dite "multi-canaux" (Laptev *et al.*, 2008 ; Wang *et al.*, 2011) est ensuite utilisée afin d'obtenir une version spatio-temporelle plus localisée du sac de mots. Elle consiste à subdiviser une vidéo selon une certaine division en espace et en temps appelée "structure". Un histogramme de mots visuels est calculé sur chaque cellule de cette structure. L'histogramme global de la vidéo est la concaténation des histogrammes de chacune de ses cellules. La subdivision de la vidéo suivant une structure particulière est appelée un "canal". La Figure 7 (Laptev *et al.*, 2008) illustre quelques exemples de structures ainsi que leurs cellules de différentes couleurs.

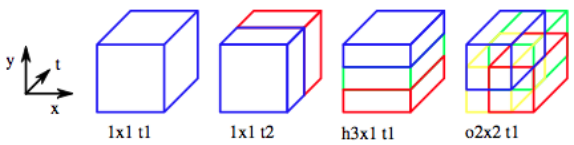

Figure 7. Exemple de différents canaux

La structure $1x1t1$ correspond à la représentation standard du sac de mot visuel. La structure $1x1t2$ correspond à une subdivision temporelle en deux cellules, tandis que la structure $o2x2t1$ correspond à une subdivision spatiale horizontale et verticale avec une zone centrale de chevauchement.

La version spatio-temporelle du sac de mots utilise différents canaux de ce type afin de combiner un plus grand nombre d'informations locales.

4.2.2. Étape de classification par SVM

En sortie de la représentation d'une séquence par l'histogramme des mots visuels extraits, une classification supervisée par SVM (Chang, Lin, 2011) est effectuée. Le SVM utilise un noyau gaussien multi-dimensionnel, permettant d'établir une distance entre des vidéos représentées par plusieurs histogrammes provenant des différents canaux (Zhang *et al.*, 2006). Le noyau est le noyau RBF qui est défini comme :

$$K_{RBF}(x_i, x_j) = exp(-\sum_{c \in C} \frac{1}{A_c} D(H_i^c, H_j^c)) \quad (1)$$

où H_i^c et H_j^c sont respectivement les histogrammes des vidéos x_i et x_j relatifs au canal c parmi l'ensemble C des canaux utilisés. $D(H_i^c, H_j^c)$ est la distance du χ^2 et A_c un coefficient de normalisation (Zhang *et al.*, 2006). Un apprentissage supervisé est réalisé pour chaque descripteur. Une fusion tardive *a posteriori* est réalisée avec la méthode Adaboost multiclasse (Hastie *et al.*, 2009) afin d'exploiter au mieux la complémentarité entre descripteurs. Des évaluations récentes ont prouvé l'efficacité de ce choix de fusion tardive dans le cadre de la reconnaissance d'actions (Peng *et al.*, 2014).

Pour les bases de données contenant une plus grande quantité de données telles que UCF-11 et UCF-50, un noyau linéaire est utilisé pour la classification (Fan *et al.*, 2008) afin de réduire les temps de calculs :

$$K_L(x_i, x_j) = (H_i^C)^T H_j^C \quad (2)$$

avec H_i^C et H_j^C qui sont les histogrammes résultant de la concaténation de tous les histogrammes des canaux de l'ensemble C.

L'approche par sac de mot visuel assure une représentation parcimonieuse des sequences vidéos. L'utilisation d'un noyau linéaire se révèle être plus efficace qu'un noyau non-linéaire lorsqu'il s'agit de discriminer, des données creuses de très grandes dimensions avec un grand nombre d'instances (Fan *et al.*, 2008). De plus, l'utilisation d'un noyau linéaire permet de génerer des dictionnaires de mots visuels de plus grande taille.

4.3. Résultats

Nous évaluons ici les performances de notre méthode sur les bases de données citées plus haut. Les taux de reconnaissance obtenus sont exposés dans le tableau 1.

4.3.1. Paramètres de la méthode

Par rapport aux méthodes présentées dans la littérature, nous avons ici peu de paramètres à fixer. Ces paramètres sont :

– C_p: Nombre de points critiques
– N: Taille des trajectoires
– S: Structure des canaux
– s: Nombre d'échelles spatio-temporelles des trajectoires multi-échelles

Nos expérimentations ont montré que les structures S et le nombre C_p de points critiques utilisés sont les paramètres qui influent le plus sur le taux de reconnaissance. La taille des trajectoires pour les bases de données traitées a été fixée à 16 frames. La Figure 8 montre l'évolution du taux de reconnaissance moyen en fonction de la variation du nombre de points critiques et du nombre de coefficients de Fourier conservés pour le descripteur FCD sur la base de données KTH. L'influence du nombre d'échelles spatio-temporelles sur la base UCF-11 est détaillée dans le tableau 2.

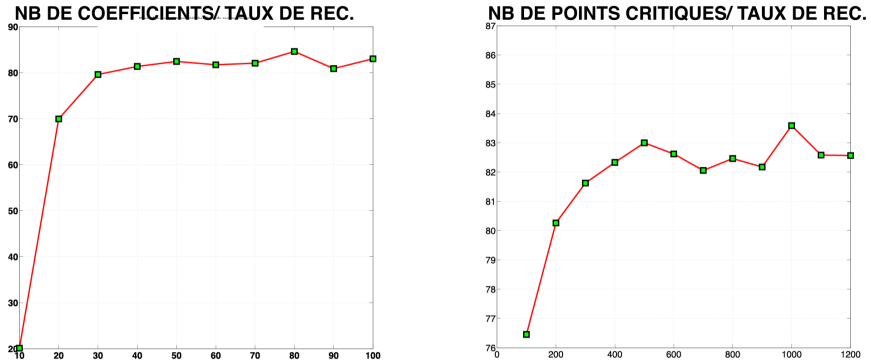

Figure 8. Nombre de coefficients de Fourier gardés et nombre de points d'intérêt obtenus en fonction du taux de reconnaissance moyen (KTH dataset). À partir d'un certain seuil, le taux de reconnaissance moyen évolue peu pour ces deux critères

4.3.2. Temps de calcul

Cette méthode a été implémentée sous Matlab sur un serveur muni d'un processeur 2 QuadCore à 3.1 Ghz et 24 GB de RAM. Nous utilisons une méthode d'estimation du flot optique avec une implémentation efficace proposée par (Sun *et al.*, 2010). Le flot optique est l'étape la plus coûteuse en termes de temps de calcul (figure 9). L'extraction des éléments d'intérêt et le calcul des trajectoires multi-échelles, qui constitue le cœur de notre méthode, sont les étapes les plus rapides.

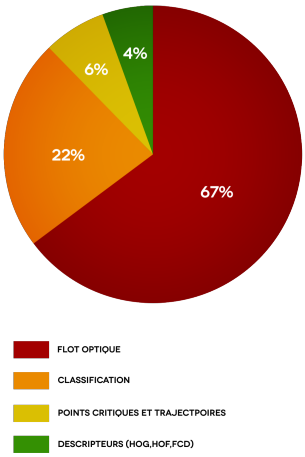

Figure 9. Proportion du temps de calcul de chaque étape de la méthode

4.3.3. Discussion des résultats

Les taux de reconnaissance des descripteurs utilisés dans notre approche sont présentés dans le tableau 1 pour chaque base de données. Les gains de reconnaissance obtenus par une fusion tardive de type Adaboost des informations de fréquence, d'orientation de mouvement, et de forme illustrent la complémentarité de ces différentes caractéristiques (3,67 % de gain en moyenne pour les bases de données traitées).

Tableau 1. Différents taux de reconnaissance obtenus sur les différentes bases de données

	KTH	Weizman	UCF-11	UCF-50
FCD	85,47 %	90,12 %	66,42 %	53,58 %
HOG	91,98 %	92,59 %	86,98 %	84,88 %
HOF	91,98 %	95,06 %	74,43 %	73,80 %
Combiné	**95,32 %**	**100 %**	**89,99 %**	**88,30 %**

L'approche multi-canaux du sac de mots visuels ainsi que le nombre de point critiques utilisés sont les paramètres qui influent le plus sur le taux de reconnaissance. Le nombre d'échelles de trajectoires participe à l'amélioration du taux de reconnaissance notamment dans le cas des vidéos génériques, où l'information fréquentielle est plus riche. On constate dans le tableau 2, pour le descripteur FCD qui encode l'information fréquentielle, que le passage d'une à trois échelles spatio-temporelles permet un gain de 3,92 %. Le descripteur HOG montre de bons résultats sur les bases de données de vidéos génériques (UCF-11, UCF-50). Le contexte spatial est très pertinent pour certaines actions, qui se déroulent dans un cadre toujours bien défini, notamment les actions d'interaction avec un objet ou les actions sportives. Le descripteur HOG en-

code l'information liée à la variation du gradient dans les vidéos, d'où la pertinence de cette caractéristique pour les vidéos génériques.

Les résultats de la méthode de compensation du mouvement de la caméra sont aussi présentés. La compensation a été effectuée dans deux situations. Le cas $s = 1$ et $s = 3$. Pour la base de données UCF-11 où les vidéos présentent différents mouvements de caméra, le gain est de 10,55 % avec $s = 1$ et de 8,83 % avec $s = 3$ pour le descripteur de trajectoires FCD. Ce résultat illustre l'intérêt de la compensation du mouvement de la caméra pour l'estimation de trajectoires. Le gain obtenu pour les descripteurs HOF et HOG montre que le flot optique obtenu après compensation est plus caractéristique du mouvement présent dans la séquence video. Les points critiques relatifs aux mouvements sont mieux localisés, et l'information encodée par HOF est à la fois moins perturbée et plus pertinente.

Dans la meilleure configuration ($s = 3$) le gain global obtenu avec la compensation du mouvement de la caméra est de 2,1 % pour la base UCF-11. Le taux de reconnaissance global de 89,99 % se situe parmi les plus élevés de la littérature pour cette base de données.

Tableau 2. Taux de reconnaissance avec variation des échelles dyadique et de la compensation du mouvement de caméra sur la base de données UCF-11.

UCF-11	$s = 1$	$s = 3$	$s = 1$ + Comp. mvt. caméra	$s = 3$ + Comp. mvt. caméra
FCD	53,50 %	57,42 %	64,05 %	66,42 %
HOG	80,34 %	84,53 %	83,34 %	86,98 %
HOF	70,06 %	74,34 %	74,06 %	74,43 %
Combiné	**82,07 %**	**86,98 %**	**87,89 %**	**89,99 %**

4.3.4. Comparaison avec les approches existantes

Notre approche est comparée avec les méthodes de la littérature dans le tableau 3 pour les différentes bases de données utilisées.

Tableau 3. Comparaison de notre approche avec la littérature

KTH	Weizman	UCF-11	UCF-50
Dollar *et al.* 89,1 %	Gorelick *et al.* 97,8 %	J. Liu *et al.* 71,2 %	Reddy *et al.* 76,90 %
Laptev *et al.* 92,1 %	Blank *et al.* 99,6 %	Wang *et al.* 85,4 %	Murthy *et al.* 87,3 %
Wang *et al.* 94,2 %	Vrigkas *et al.* 100 %	Reddy *et al.* 87,1 %	Wang *et al.* 91,2 %
Vrigkas *et al.* **98,3 %**	Gorelick *et al.* 100 %	**Vrigkas** *et al.* **95,1 %**	**Peng** *et al.* **92,3 %**
Notre méthode 95,32 %	**Notre méthode 100 %**	Notre méthode 89,99 %	Notre méthode 88,30 %

Le tableau 4 montre le nombre moyen de caractéristiques générées par notre méthode ainsi que celles de (Wang *et al.*, 2013) et (Shi *et al.*, 2013). Cela donne une indication sur le nombre de caractéristiques à obtenir pour atteindre un taux de reconnaissance donné.

Tableau 4. Caractéristiques moyennes extraites par image. Les chiffres sont relatifs à UCF-50

Méthode	(Wang *et al*., 2013)	(Shi *et al*., 2013)	**Notre Approche**
caract./frame	230	65.3	**70,6**
%	91,2 %	83,3 %	**88,3 %**

(Shi *et al*., 2013) proposent une sélection aléatoire de 10 000 caractéristiques dans un ensemble extrait sur une grille dense. (Wang *et al*., 2013) produisent un taux de caractéristiques par image proche de 230 pour le traitement de vidéos en conditions non contrôlées. Cette méthode donne un taux de reconnaissance parmi les meilleurs de l'état de l'art mais génère un nombre de caractéristiques très élevé, en plus de l'utilisation de 8 échelles spatiales de trajectoires et de 30 canaux de partition spatio-temporelle et 15 % du temps d'execution de la méthode est réservé à l'écriture des données générées sur disque dur. (Murthy, Goecke, 2013) ont testé sur une vidéo le nombre de caractéristiques mis en œuvre comparé à la méthode de (Wang, Schmid, 2013) après l'étape de correspondance de trajectoires. Dans le cas de l'utilisation d'un seul canal et de trajectoires de taille 15, la méthode de correspondance de trajectoires met en oeuvre 1,85 fois moins de trajectoires que la méthode des trajectoires denses (11 657 contre 21 647 caractéristiques). En se reportant au taux de caractéristiques par image de (Wang, Schmid, 2013), cela donnerait un taux moyen de 124,32 pour (Murthy, Goecke, 2013) sur la base UCF-50 avec un taux de 87,3 %.

Pour la base UCF-50, notre méthode met en oeuvre 1 200 points critiques par échelles et par canaux, ce qui nous donne un total de 10 800 points par vidéo et un taux de caractériques par image moyen de 70,6. On constate ici que la différence de complexité des meilleurs approches de la littérature avec la notre permet de relativiser l'écart de taux de reconnaissance obtenu sur cette base de données (88,30 %).

5. Conclusion

Ce papier présente une méthode originale de reconnaissance d'actions humaines dans des vidéos. Les vidéos sont caractérisées par des points critiques calculés à partir du flot optique, ainsi que par les trajectoires de ces points à différentes échelles spatio-temporelles.

La caractérisation fréquentielle des trajectoires de mouvements, combinée à l'orien tation locale du flot optique et à l'information de contours, permet d'atteindre des résultats parmi les plus élevés de la littérature. Le fait de ne signer que le mouvement des points critiques est un avantage non négligeable en termes de complexité. En effet, on atteint des taux de reconnaissance proches des méthodes denses, en calculant beaucoup moins de caractéristiques que ces dernières. La forte cohérence des points critiques face aux mouvement présents ainsi que la pertinence des éléments fusionnés témoignent de ces résultats.

Les taux de reconnaissance obtenus sur les différentes bases de données traitées attestent de la performance de la méthode dans différentes situations : reconnaissance d'actions de videos dans des conditions contrôlées (KTH) ou avec peu de contraintes d'acquisitions (UCF-11), discrimination d'actions différentes mais visuellement similaires (Weizmann), ou bien discrimination d'un grand nombre de classes d'actions (UCF-50).

De par les résultats obtenus sur des bases de données d'actions humaines de référence, nos perspectives se portent vers la reconnaissance et la caractérisation d'actions dites complexes, ou activités, lorsque ces dernières sont représentées comme un enchainement séquentiel au cours du temps d'actions élémentaires.

Notre méthode sera aussi évaluée pour la reconnaissance de textures dynamiques (Péteri *et al*., 2010). Nous pensons que l'utilisation de points critiques du flot optique ainsi que l'information fréquentielle peut être pertinente dans ce cadre.

Bibliographie

Chang C.-C., Lin C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, p. 27:1–27:27.

Dollar P., Rabaud V., Cottrell G., Belongie S. (2005, oct.). Behavior recognition via sparse spatio-temporal features. In *Ieee international workshop on visual surveillance and performance evaluation of tracking and surveillance*, p. 65 - 72.

Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, vol. 9, p. 1871–1874.

Gorelick L., Blank M., Shechtman E., Irani M., Basri R. (2007, December). Actions as space-time shapes. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, n° 12, p. 2247–2253.

Hastie T., Rosset S., Zhu J., Zou H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, vol. 2, n° 3, p. 349–360.

Jain M., Jégou H., Bouthemy P. (2013, avril). Better exploiting motion for better action recognition. In *Proc. conf. comp. vision pattern rec.* Portland, États-Unis.

Laptev I. (2005). On space-time interest points. *Int. J. Computer Vision*, vol. 64, n° 2-3, p. 107–123.

Laptev I., Marszalek M., Schmid C., Rozenfeld B. (2008, June). Learning realistic human actions from movies. In *Proc. conf. comp. vision pattern rec.*, p. 1 -8.

Lazebnik S., Schmid C., Ponce J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. conf. comp. vision pattern rec.*, vol. 2, p. 2169-2178.

Liu J., Luo J., Shah M. (2009, juin). Recognizing realistic actions from videos "in the wild". In *Proc. conf. comp. vision pattern rec.*, p. 1996–2003.

Murthy O., Goecke R. (2013, December). Ordered trajectories for large scale human action recognition. In *Proc. int. conf. computer vision*, p. 412-419.

Peng X., Wang L., Wang X., Qiao Y. (2014). Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computing Research Repository*, vol. abs/1405.4506.

Péteri R., Fazekas S., Huiskes M. J. (2010). DynTex : a Comprehensive Database of Dynamic Textures. *Pattern Recognition Letters*.

Raptis M., Soatto S. (2010). Tracklet descriptors for action modeling and video analysis. In *Proc. europ. conf. computer vision*, p. 577–590. Berlin, Heidelberg, Springer-Verlag.

Reddy K. K., Shah M. (2012). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, vol. 24, n° 5, p. 971–981.

Schuldt C., Laptev I., Caputo B. (2004). Recognizing human actions: a local svm approach. In *Proc. int. conf. pattern recognition*, vol. 3, p. 32-36 Vol.3.

Shi F., Petriu E., Laganiere R. (2013, June). Sampling strategies for real-time action recognition. In *Proc. conf. comp. vision pattern rec.*

Sun D., Roth S., Black M. (2010). Secrets of optical flow estimation and their principles. In *Proc. conf. comp. vision pattern rec.*, p. 2432-2439.

Ullah M. M., Laptev I. (2012). Actlets: A novel local representation for human action recognition in video. In *Proc. ieee international conference on image processing*, p. 777-780.

Vrigkas M., Karavasilis V., Nikou C., Kakadiaris A. (2014). Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding*, vol. 119, p. 27 – 40.

Wang H., Klaser A., Schmid C., Liu C.-L. (2011, June). Action recognition by dense trajectories. In *Proc. conf. comp. vision pattern rec.*, p. 3169 -3176.

Wang H., Kläser A., Schmid C., Liu C.-L. (2013, mai). Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Computer Vision*, vol. 103, n° 1, p. 60–79.

Wang H., Muneeb Ullah M., Kläser A., Laptev I., Schmid C. (2009). Evaluation of local spatio-temporal features for action recognition. In *University of central florida, u.s.a.*

Wang H., Schmid C. (2013). Action Recognition with Improved Trajectories. In *Proc. int. conf. computer vision*, p. 3551-3558. Sydney, Australie, IEEE.

Willems G., Tuytelaars T., Gool L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. europ. conf. computer vision*, p. 650–663. Berlin, Heidelberg, Springer-Verlag.

Zhang J., Marszalek M., Lazebnik S., Schmid C. (2006). Local features and kernels for classification of texture and object categories: A comprehensive study. In *Proc. conf. comp. vision pattern rec.*, p. 13-13.

Article soumis le 17/12/2014
Accepté le 10/06/2015