Reconnaissance d'Actions par Modélisation du Mouvement

Yassine Benabbas*, Adel Lablack*, Thierry Urruty*, Chabane Djeraba*

*LIFL UMR CNRS 8022, Université de Lille1, TELECOM Lille1
IRCICA, Parc de la Haute Borne, 56950 Villeneuve d'Ascq
{yassine.benabbas,adel.lablack,thierry.urruty,chabane.djeraba}@lifl.fr

Résumé. Cet article propose une approche utilisant les modèles de direction et de magnitude de mouvement pour détecter les actions qui sont effectuées par des êtres humains dans des séquences vidéo. Des mélanges Gaussiens et de lois de von Mises sont estimés à partir des orientations et des magnitudes des vecteurs du flux optique calculés pour chaque bloc de la scène. Les paramètres de ces modèles sont estimés grâce à un algorithme d'apprentissage en ligne. Les actions sont reconnues grâce à une mesure qui se base sur la distance de Bhattacharyya et qui permet de comparer le modèle d'une séquence donnée avec les modèles créés à partir de séquences d'apprentissage. L'approche proposée est évaluée sur deux ensembles de vidéos contenant des actions variées exécutées aussi bien dans des environnements intérieur qu'extérieur.

1 Introduction

La reconnaissance des actions est un sujet particulièrement complexe dans le domaine de la vision par ordinateur. Cela consiste en la classification automatique des actions ou des activités réalisées par un individu dans une séquence vidéo. La reconnaissance des actions est cruciale dans de nombreux domaines comme la vidéo-surveillance, l'interaction homme-machine et l'indexation des vidéos.

L'objectif d'un système de reconnaissance d'actions est de reconnaitre des actions simples de la vie courante dans une vidéo (ex : marcher, répondre au téléphone, sauter...) à partir de vidéos de référence. Ces actions répondent à des modèles de mouvements simples effectués par une seule et même personne durant un laps de temps court. Certaines approches détectent les actions à partir d'images fixes, tandis que d'autres ont recours à des vidéos stéréoscopiques ou à des maillages 3D (Ganesh et Bajcsy, 2008). Dans cet article, nous traitons les séquences vidéo enregistrées par des caméras monoculaires car elles permettent de détecter des actions en combinant des informations spatiales et temporelles (Johansson et al., 1994). Cet intérêt pour les vidéos monoculaires résulte du fait qu'elles sont couramment utilisées, moins gourmandes en ressources et plus économiques.

Cet article présente une méthodologie qui permet de reconnaître des actions en se basant sur l'analyse du mouvement des sujets. Il est organisé comme suit, les travaux antérieurs sont passés en revue dans la Section 2. Nous décrivons ensuite notre approche dans la Section 3 en détaillant ses deux phases principales qui sont la *construction de modèles* et la *reconnaissance*

d'action. L'approche proposée est évaluée sur deux ensembles de données et les résultats expérimentaux sont rapportés dans la Section 4. Nous concluons et proposons plusieurs pistes pour nos travaux futurs dans la Section 5.

2 Travaux antérieurs

Durant ces dernières années, de nombreuses approches de reconnaissance des actions ont été proposées. Elles sont décrites dans des études bibliographiques (Poppe, 2010; Turaga et al., 2008). Ces techniques ont été classées en fonction de la méthode de représentation des images et de l'algorithme pour la classification de l'action comme suit :

Représentation des images : le calcul des caractéristiques à partir des images de la vidéo tient compte de la dimension temporelle. Il s'agit généralement des vecteurs de flux optique (Ali et Shah, 2010), de caractéristiques spatio-temporelles comme les cuboïdes (Dollar et al., 2005) ou des caractéristiques hessiennes (Willems et al., 2008). Un descripteur est ensuite créé pour représenter la séquence vidéo. (Fathi et Mori, 2008) calculent les descripteurs par apprentissage de classificateurs Ada-boost à partir des caractéristiques de bas niveau, alors que (Messing et al., 2009) calculent la trajectoire des points en mouvement. Certains descripteurs tels que HOG/HOF (Laptev et al., 2008), HOG3D (Kläser et al., 2008) et ESURF (SURF étendu) (Willems et al., 2008) sont basés sur l'analyse spatio-temporelle locale des points en mouvement. Les meilleures méthodes de représentation des images sont celles qui discernent efficacement les actions en classes différentes et qui s'exécutent en temps réel.

Classification de l'action : c'est le mécanisme qui permet de classifier une action. Elle peut être effectuée en utilisant un classificateur comme SVM (Mauthner et al., 2009), SOM (Self-Organizing Map) (Huang et Wu, 2009), un processus Gaussien (Wang et al., 2009b), une fonction de distance (Yang et al., 2009), ou un modèle discriminant tel que HCRF (Hidden Conditional Random Field ; Champ Aléatoire Conditionnel Caché) (Zhang et Gong, 2010).

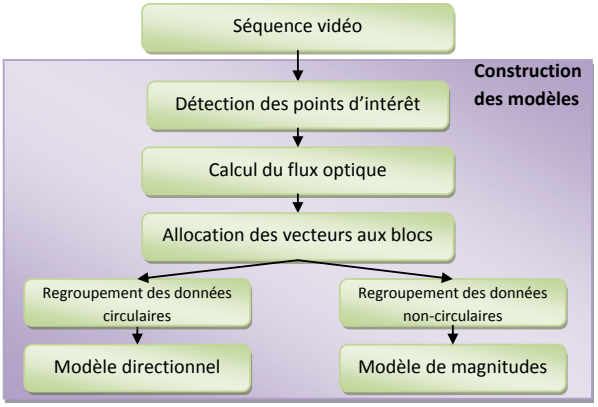
Afin d'effectuer des tests ou comparer différentes approches, les bases de vidéos telles que KTH (Laptev et Lindeberg, 2004) et ADL (Activities of Daily Living ; Activités de la Vie Quotidienne) (Messing et al., 2009) sont utilisées.

Les descripteurs spatio-temporels locaux ont récemment vu leur popularité se développer et se sont avérés efficaces dans le cadre de la reconnaissance des actions humaines (Wang et al., 2009a). Nos modèles qui sont inspirés par les descripteurs HOG/HOF proposés par (Laptev et al., 2008) permettent d'extraire les principales orientations/magnitudes du mouvement et leur attribuent une variance et un poids au lieu d'histogrammes calculés à partir de la fréquence d'observation des vecteur de mouvement. Nous proposons une approche dont l'originalité repose sur l'utilisation de modèles de direction et de magnitude pour représenter des actions sans passer par la détection des membres du corps humain. En effet, elle s'appuie sur les vecteurs de flux optique en tant que caractéristiques permettant de construire le modèle associé à une séquence qui est estimé et mis à jour en temps réel. Notre approche extrait les magnitudes et orientations de mouvement principales dans chaque bloc de la scène. Nous avons choisi une représentation dense car ce type d'échantillonnage offre de meilleurs résultats (Wang et al., 2009a). Les actions sont ensuite détectées par le biais d'une mesure de distance appliquée entre le modèle associé à une séquence de référence et celui associé à une séquence requête. Dans ce qui suit, une séquence requête est une séquence (ou vidéo) dont on cherche à reconnaitre l'action.

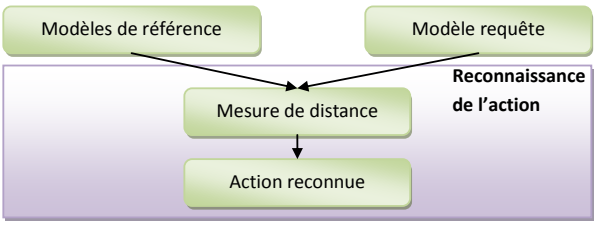
3 Description de l'approche

Pour détecter les actions réalisées par une seule personne, nous proposons une approche dont les principales étapes sont illustrées dans la Figure 1. Ces étapes sont divisées en deux phases principales :

– Construction des modèles : elle permet de quantifier le mouvement à partir des vecteurs de flux optique afin d'estimer le modèle directionnel et le modèle de magnitude pour l'intégralité de la séquence.
– Reconnaissance de l'action : elle permet de reconnaitre l'action dans une vidéo en comparant son modèle avec les modèles des séquences vidéo de référence par le biais d'une mesure de distance.



(a)



(b)

FIG. 1 – *Etapes de l'approche. (a) Phase de construction des modèles. (b) Phase de reconnaissance de l'action.*

3.1 Construction des modèles

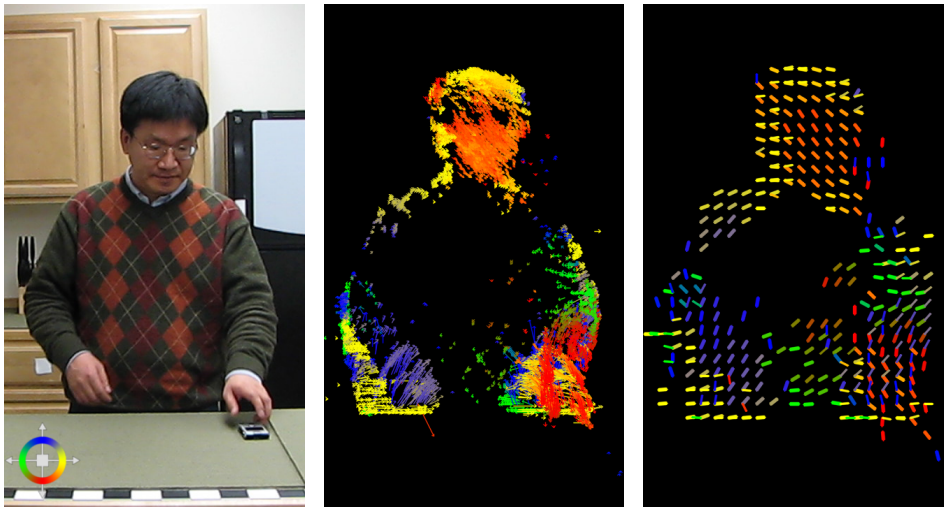
Pour construire le modèle d'une séquence vidéo, nous commençons par extraire un ensemble de points d'intérêt dans chaque image. Nous avons utilisé le détecteur de points d'intérêt de Shi et Tomasi (Shi et Tomasi, 1994) qui permet de trouver les coins possédant une valeur propre élevée. Nous considérons également que, dans les vidéos traitées, la position de la ca-

méra et les conditions d'éclairage permettent d'obtenir un grand nombre de points d'intérêt pouvant être facilement détectés et suivis.

Après avoir défini l'ensemble des points d'intérêt, nous suivons leurs déplacements sur les images suivantes grâce aux vecteurs de flux optique. Pour cela, nous utilisons l'implémentation de Bouguet (Bouguet, 2000) de l'algorithme de suivi KLT (Lucas et Kanade, 1981) qui s'avère rapide et efficace pour gérer les points se trouvant à proximité du bord de l'image. Le résultat est un ensemble de vecteurs de mouvement où un vecteur est défini par une origine, une orientation et une magnitude.

L'étape suivante consiste à diviser la scène en une grille de $M \times N$ blocs. Puis, chaque vecteur de mouvement est associé au bloc qui lui correspond, selon son origine. La taille des blocs influe sur la précision du système et sera étudiée dans la Section 4.3.

Un algorithme de regroupement de données circulaires est ensuite appliqué aux orientations des vecteurs de flux optique dans chaque bloc. L'ensemble des $M \times N$ distributions circulaires associées est appelé "modèle directionnel". La Figure 2 montre la construction d'un modèle directionnel associé à l'action 'answerPhone'.



(a) (b) (c)

FIG. 2 – *Modèle directionnel pour l'action 'answerPhone'. (a) image courante, (b) vecteurs de flux optique, (c) modèle directionnel associé à la séquence vidéo.*

Dans cet article, nous regroupons les données circulaires en utilisant un mélange de lois von Mises. Ainsi, la probabilité d'obtenir une orientation θ par rapport à un bloc $B_{x,y}$ est définie par la formule suivante :

$$p_{x,y}(\theta) = \sum_{i=1}^{K} \psi_{i_{x,y}} \cdot V\left(\theta; \phi_{i_{x,y}}, \gamma_{i_{x,y}}\right) \quad (1)$$

où K représente le nombre de lois du mélange. Nous avons choisi empiriquement $K = 4$, pour correspondre aux quatre points cardinaux. $\psi_{i_{x,y}}$, $\phi_{i_{x,y}}$, $\gamma_{i_{x,y}}$ sont respectivement le poids,

l'angle moyen et le paramètre de concentration de la $i^{\text{ème}}$ distribution du bloc $B_{x,y}$. $V(\theta;\phi,\gamma)$ est la loi de von Mises de direction ϕ avec un paramètre de concentration γ. Elle possède la fonction de densité de probabilité suivante sur l'intervalle $[0, 2\pi[$:

$$V(\theta;\phi,\gamma) = \frac{1}{2\pi I_0(\gamma)} \exp\left[\gamma \cos(\theta - \phi)\right] \tag{2}$$

où $I_0(\gamma)$ est la fonction de Bessel modifiée de première espèce d'ordre 0 définie par :

$$I_0(\gamma) = \sum_{r=0}^{\infty} \left(\frac{1}{r!}\right)^2 \left(\frac{1}{2}\gamma\right)^{2r} \tag{3}$$

Par analogie, nous regroupons les magnitudes des vecteurs du flux optique dans chaque bloc grâce à des mélanges Gaussiens. L'ensemble des mélanges Gaussiens estimés représente le modèle de magnitude. Ainsi, la probabilité d'une magnitude v par rapport au bloc $B_{x,y}$ est définie de la façon suivante :

$$p_{x,y}(v) = \sum_{i=1}^{J} \omega_{i_{x,y}} G(v; \mu_{i_{x,y}}, \sigma^2_{i_{x,y}}) \tag{4}$$

où $\omega_{i_{x,y}}, \mu_{i_{x,y}}, \sigma^2_{i_{x,y}}$ sont respectivement le poids, la moyenne et la variance de la $i^{\text{ème}}$ Gaussienne. J est le nombre de Gaussiennes ($J = 4$ dans nos expérimentations).

Pour chaque image, nous mettons à jour les paramètres des mélanges Gaussiens grâce à une approximation de k-means décrite dans (Kaewtrakulpong et Bowden, 2001). Nous l'utilisons également pour estimer les paramètres des mélange de lois de von Mises en adaptant l'algorithme afin de gérer les données circulaires et en prenant en compte l'inverse de la variance en tant que paramètre de dispersion, $\gamma = 1/\sigma^2$.

Nous annotons ci-dessous le modèle de la séquence s par $Sm(s) = (Dm(s), Mm(s))$, où $Dm(s)$ et $Mm(s)$ sont respectivement le modèle directionnel et le modèle de magnitude associés à la séquence s. La Figure 3 montre les modèles directionnels et de magnitude de quelques séquences vidéo issues de la base KTH.

3.2 Reconnaissance de l'action

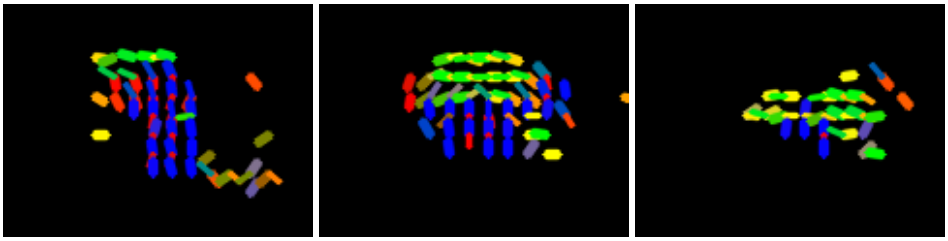
Une fois le modèle de la séquence vidéo calculé, nous détectons l'action correspondant à cette séquence requête en fonction des vidéos de référence. Les actions sont détectées en comparant le modèle d'une séquence requête avec les modèles associés aux séquences de référence en utilisant une mesure de distance. L'action associée au modèle ayant la distance la plus petite par rapport au modèle d'une séquence requête est retenue.

Soient $T = \{t_1, t_2, ..., t_n\}$ un ensemble de n séquences avec leurs modèles respectifs $\{Sm(t_1), Sm(t_2), ..., Sm(t_n)\}$ et q une séquence requête avec son modèle $Sm(q)$. La distance entre $Sm(q)$ et une séquence de référence $Sm(t_l)$ est définie par :

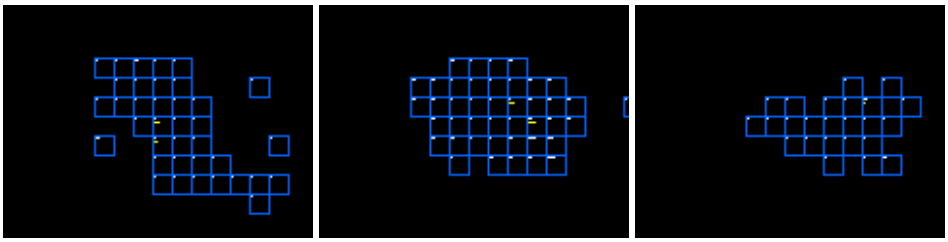
$$D(Sm(q), Sm(t_l)) = Norm(A_{Dm(q),Dm(t_l)}) + Norm(B_{Mm(q),Mm(t_l)}) \tag{5}$$

où $Norm$ correspond à la norme L2. Les matrices $A_{Dm(q),Dm(t_l)}$ et $B_{Mm(q),Mm(t_l)}$ de $W \times H$ contiennent les distances entre chaque élément des deux modèles directionnel

(a) Echantillon d'images



(b) Modèle directionnel



(c) Modèle de magnitude

FIG. 3 – *Échantillon d'images avec les modèles de direction et de magnitude qui leur sont associés*

$Dm(q)$ et $Dm(t_l)$ et des deux modèles de magnitude $Mm(q)$ et $Mm(t_l)$. Chaque élément $A_{M,M'}(x,y)$ est défini par la formule suivante :

$$A_{M,M'}(x,y) = \sum_{i=1}^{K} \left(\psi_{i_{x,y}} \psi'_{i_{x,y}} Dist_d(V_{i_{x,y}}, V'_{i_{x,y}}) \right) \tag{6}$$

où $\psi_{i_{x,y}}$ (resp. $\psi'_{i_{x,y}}$) et $V_{i_{x,y}}$ (resp. $V'_{i_{x,y}}$) représentent le poids et la variance de la $i^{\text{ème}}$ loi de von Mises associés au modèle directionnel M (resp. M') dans le bloc $B_{x,y}$. $Dist_d(V,V')$ est la distance de Bhattacharyya entre les deux lois de von Mises V et V' définie par l'équation suivante :

$$Dist_d(V,V') = \sqrt{1 - \int_{-\infty}^{+\infty} \sqrt{V(\theta)V'(\theta)} d\theta} \tag{7}$$

où $Dist_d(V, V')$ est comprise entre 0 et 1. Cette équation peut être calculée grâce à cette solution de forme fermée :

$$Dist_d(V, V') = \sqrt{1 - \sqrt{\frac{1}{I_0(\gamma)I_0(\gamma')}} I_0\left(\frac{\sqrt{\gamma^2 + \gamma'^2 + 2\gamma\gamma' cos(\phi - \phi')}}{2}\right)} \quad (8)$$

où ϕ (resp. ϕ') et γ (resp. γ') sont respectivement l'angle moyen et le paramètre de dispersion de la distribution V (resp. V'). Une autre mesure de distance est étudiée dans (Benabbas et al., 2010)

Par analogie, nous définissons chaque élément $B_{N,N'}(x, y)$ par l'équation suivante :

$$B_{N,N'}(x, y) = \sum_{i=1}^{K} \left(\omega_{i_{x,y}} \omega'_{i_{x,y}} Dist_m(G_{i_{x,y}}, G'_{i_{x,y}})\right) \quad (9)$$

où $\omega_{i_{x,y}}$ (resp. $\omega'_{i_{x,y}}$) et $G_{i_{x,y}}$ (resp. $G'_{i_{x,y}}$) représentent le poids de la $i^{\text{ème}}$ Gaussienne associée au modèle de magnitude N (resp. N') dans le bloc $B_{x,y}$. $Dist_m(G, G')$ est la distance de Bhattacharyya entre deux Gaussiennes G et G' définies dans la solution de forme fermée suivante :

$$Dist_m(G, G') = \frac{(\mu - \mu')^2}{4(\sigma^2 + \sigma'^2)} + \frac{1}{2} \ln\left(\frac{\sigma^2 + \sigma'^2}{2\sigma\sigma'}\right) \quad (10)$$

où μ (resp. μ') et σ^2 (resp. σ'^2) sont respectivement la moyenne et la variance de la Gaussienne G (resp. G').

4 Expérimentations et résultats

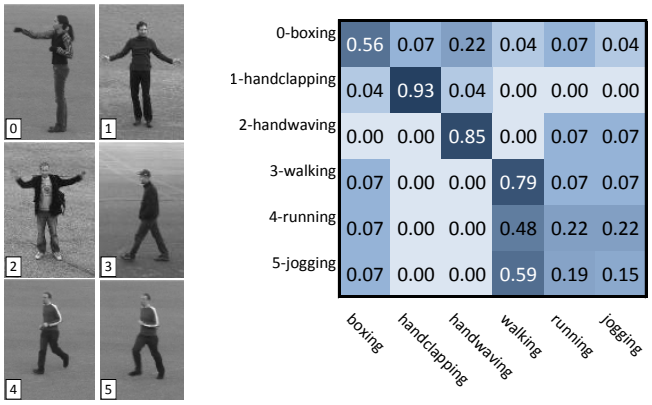
Nous démontrons dans cette section l'efficacité de notre approche à l'aide de deux ensembles de vidéos portant sur une variété d'actions quotidiennes. Les matrices de confusion sont présentées suivies des effets de la taille des blocs et du nombre de classes d'actions sur les performances du système.

4.1 Efficacité de la reconnaissance des actions

Base vidéo KTH (Laptev et Lindeberg, 2004) : c'est une base de vidéos de faible résolution (images en niveau de gris de 160×120 pixels) regroupant 6 actions effectuées plusieurs fois par 25 personnes. Cette base contient des vidéos en environnement intérieur en extérieur et les personnes portent des tenues vestimentaires différentes. Nous divisons l'ensemble de données en deux ensembles, comme suggéré par Schuldt et al. (Schuldt et al., 2004) : un ensemble d'apprentissage qui contient les séquences de référence (16 personnes) et un ensemble de test qui contient les séquences requête (9 personnes). L'ensemble d'apprentissage comporte les personnes 'person01' à 'person16' et l'ensemble de test comporte les personnes 'person17' à 'person25'. Nous utilisons des blocs de taille de 5×5.

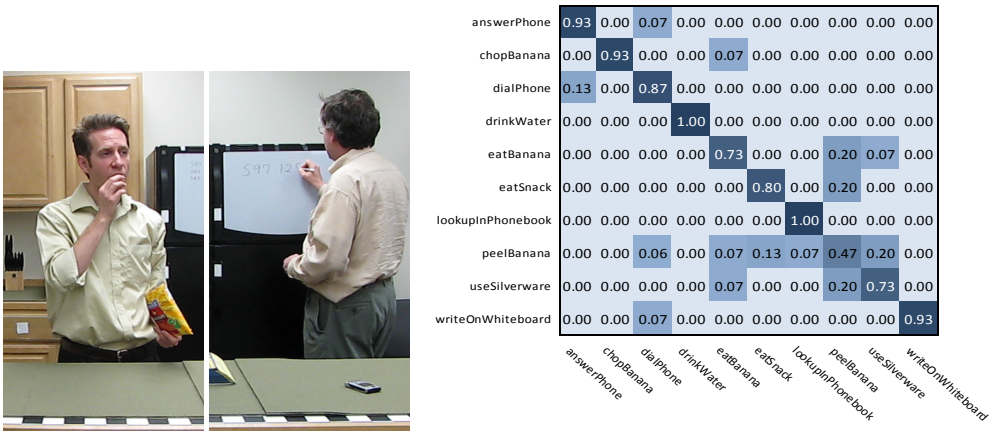
Quelques exemples d'actions ainsi que la matrice de confusion sont indiqués dans la Figure 4. Notre approche aboutit à des résultats satisfaisants pour les trois premières actions de

la base de vidéos lorsque la personne est immobile. En revanche, notre système assimile les actions 'run' et 'jogging' à l'action 'walk'. Ceci est dû au fait que ces actions diffèrent légèrement de par la vitesse et la longueur de la foulée tout en ayant une orientation similaire.



(a) (b)

FIG. 4 – *Résultats de l'ensemble de données KTH. (a) Échantillon d'actions, (b) Matrice de confusion utilisant un bloc de* 5×5.



(a) (b)

FIG. 5 – *Résultats pour la base ADL. (a) Échantillons d'actions. (b) Matrice de confusion utilisant des blocs de* 5×5 *pixels.*

Base vidéo ADL (Activities of Daily Living - Activités de la Vie Quotidienne) (Messing et al., 2009) : c'est une base de vidéos haute définition (1280×720 pixels) regroupant 10 actions courantes du quotidien (ex : peelBanana, useSilverware, answerPhone) effectuées par

5 personnes différentes. Nous suivons le protocole "leave-one-out" dans notre expérimentation. Pour cela, nous prenons en compte une séquence en tant que séquence requête, et toutes les autres comme séquences de référence pour la phase de reconnaissance d'action. Cette procédure est effectuée pour toutes les séquences, et la moyenne des résultats est calculée pour chaque classe d'action.

Dans la Figure 5, nous présentons la matrice de confusion obtenue dans le cadre de notre approche avec cette base de vidéos. L'approche obtient une précision moyenne de 0.84 pour des blocs de taille 5×5 pixels. Ce résultat est très satisfaisant, toutefois, l'action "peelBanana" peut être confondue avec les actions "eatSack" et "useSilverware" car elles ont un comportement initial similaire qui consiste à ramener un objet depuis le potager.

4.2 Étude comparative

Nous comparons notre approches avec d'autres en utilisant les bases de vidéos KTH et ADL, et présentons leurs précisions dans le tableau 1.

Méthode	ADL	KTH
Approche proposée	0.84	0.58
Historique des vitesses (Messing et al., 2009)	0.63	0.74
Points d'intérêt spatio-temporels (Laptev et al., 2008)	0.59	0.80
Cuboïdes spatio-temporels (Dollar et al., 2005)	0.36	0.66

TAB. 1 – *Comparaison pour 2 bases de vidéos.*

Il s'avère que les approches basées sur les descripteurs spatio-temporels locaux (Dollar et al., 2005; Laptev et al., 2008) et l'historique des vitesses (Messing et al., 2009) aboutissent à de meilleurs résultats que notre système pour la base KTH. Ce dernier a recours à la vitesse des points d'intérêt en tant que descripteurs de bas niveau. Cependant, notre système est plus performant avec la base ADL car il combine à la fois les informations relatives à la magnitude du mouvement et à l'orientation.

Par rapport aux caractéristiques HOG/HOF (Laptev et al., 2008), notre modèle de scène assimile les principales orientations et magnitudes, et il ne prend pas en compte les mouvements soumis au bruit. De plus, chaque mélange de lois renvoie des orientations moyennes avec les variances et poids correspondants, tandis que les descripteurs HOG/HOF calculent des histogrammes sur des gradients orientés (HOG) et des flux optiques (HOF) moins précis. Notre approche est notamment efficace lors de l'utilisation de la base de vidéos de haute résolution ADL car elle s'appuie sur l'information de mouvement qui est plus exacte. Néanmoins, elle souffre d'un manque de précision sur les vidéos basse résolution de la base KTH.

4.3 Etude du nombre de classes d'actions et de la taille des blocs

Nous étudions l'influence de la taille des blocs et le nombre de classes d'action avec l'ensemble KTH. Ainsi, nous avons répété l'expérience pour chaque élément de l'ensemble des sous-ensembles des actions de la base KTH pour les actions suivantes : handshaking,

boxing, handwaving, walking, running et jogging. Nous avons respectivement noté ces actions $A = \{0, 1, 2, 3, 4, 5\}$. Les graphes de la figure 6 montrent la précision de notre système pour chaque sous-ensemble de A. Le graphe bleu est obtenu pour des blocs de taille 5×5 pixels, tandis que le graphe rouge correspond à des blocs de taille 10×10 pixels.

Le taux de précision le plus bas ($\sim 40\%$) est atteint lorsque les actions 345 sont combinées (correspondant aux actions trotter, courir et marcher). Cela souligne la difficulté à différencier la vitesse de chaque action dans le cadre de vidéos basse résolution.

Nos expériences montrent également que le fait d'augmenter la taille des blocs réduit la précision globale du système. Cependant, le temps de traitement est lui aussi diminué. Par ailleurs, l'augmentation du nombre de séquences de référence allonge la durée du traitement.

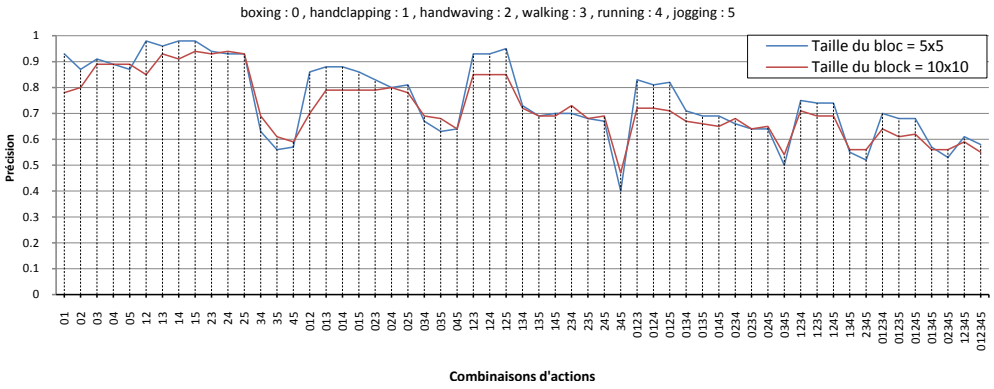


FIG. 6 – *Influence de la taille des blocs et de la combinaison des actions sur la précision.*

5 Conclusion

Nous avons présenté un système de reconnaissance d'actions performant qui se base sur les modèles de direction et les modèles de magnitude du mouvement. Nous avons extrait les vecteurs de flux optique des séquences vidéo pour acquérir des modèles statistiques sur l'orientation et la magnitude du mouvement. Le résultat est un modèle de séquence vidéo qui estime les principales orientations et magnitudes dans tous les blocs de la scène. Nous avons utilisé une mesure de distance pour détecter une action en comparant le modèle d'une séquence requête à des modèles de référence. En s'appuyant sur l'orientation et la magnitude du mouvement, notre approche aboutit à des résultats prometteurs comparés à d'autres approches de l'état de l'art, notamment sur des vidéos haute définition. Ainsi, nos travaux futurs s'orienteront vers l'amélioration de la flexibilité de notre approche par rapport à l'ajout ou la suppression de classes d'action et la reconnaissance d'actions dans des applications en temps réel.

Remerciements : Ce projet est soutenu par le projet européen MIDAS (Multimodal Interfaces for Disabled and Ageing Society (MIDAS) ITEA 2-07008) et le projet ANR CAnADA (Comportements Anormaux Analyse Détection Alerte).

Références

Ali, S. et M. Shah (2010). Human action recognition in videos using kinematic features and multipleinstance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 32*(2), 288–303.

Benabbas, Y., A. Lablack, N. Ihaddadene, et C. Djeraba (2010). Action recognition using direction models of motion. In *International Conference on Pattern Recognition (ICPR)*.

Bouguet, J.-Y. (2000). Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. Intel Corporation Microprocessor Research Labs.

Dollar, P., V. Rabaud, G. Cottrell, et S. Belongie (2005). Behavior recognition via sparse spatio-temporal features. In *2nd International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (PETS)*, pp. 65–72.

Fathi, A. et G. Mori (2008). Action recognition by learning mid-level motion features. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ganesh, S. et R. Bajcsy (2008). Recognition of human actions using an optimal control based motor model. In *Workshop on Applications of Computer Vision (WACV)*, pp. 1 –6.

Huang, W. et J. Wu (2009). Human action recognition using recursive self organizing map and longest common subsequence matching. In *International Workshop on Applications of Computer Vision (WACV)*, pp. 1 –6.

Johansson, G., S. S. Bergstrom, W. Epstein, et G. Jansson (1994). *Perceiving Events and Objects*. Lawrence Erlbaum Associates.

Kaewtrakulpong, P. et R. Bowden (2001). An improved adaptive background mixture model for realtime tracking with shadow detection. In *2nd European Workshop on Advanced Video Based Surveillance Systems (AVBS)*.

Kläser, A., M. Marszałek, et C. Schmid (2008). A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference (BMVC)*.

Laptev, I. et T. Lindeberg (2004). Velocity adaptation of space-time interest points. In *International Conference on Pattern Recognition (ICPR)*, pp. 52–56.

Laptev, I., M. Marszałek, C. Schmid, et B. Rozenfeld (2008). Learning realistic human actions from movies. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lucas, B. et T. Kanade (1981). An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 674–679.

Mauthner, T., P. M. Roth, et H. Bischof (2009). Instant action recognition. In *16th Scandinavian Conference on Image Analysis (SCIA)*.

Messing, R., C. Pal, et H. Kautz (2009). Activity recognition using the velocity histories of tracked keypoints. In *International Conference on Computer Vision (ICCV)*.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing (IVC) 28*(6), 976 – 990.

Schuldt, C., I. Laptev, et B. Caputo (2004). Recognizing human actions : A local svm approach. In *International Conference on Pattern Recognition (ICPR)*.

Shi, J. et C. Tomasi (1994). Good features to track. In *Internatioal Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 593–600.

Turaga, P., R. Chellappa, V. S. Subrahmanian, et O. Udrea (2008). Machine recognition of human activities : A survey. *IEEE Transactions on Circuits and Systems for Video Technology 18*(11), 1473–1488.

Wang, H., M. M. Ullah, A. Kläser, I. Laptev, et C. Schmid (2009a). Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference (BMVC)*, pp. 127.

Wang, L., H. Zhou, S.-C. Low, et C. Leckie (2009b). Action recognition via multi-feature fusion and gaussian process classification. In *International Workshop on Applications of Computer Vision (WACV)*.

Willems, G., T. Tuytelaars, , et L. V. Gool (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision (ECCV)*.

Yang, W., Y. Wang, et G. Mori (2009). Efficient human action detection using a transferable distance function. In *Asian Conference on Computer Vision (ACCV)*.

Zhang, J. et S. Gong (2010). Action categorization with modified hidden conditional random field. *Pattern Recognition (PR) 43*(1), 197–203.

Summary

This paper proposes an approach that uses direction and magnitude models to perform human action recognition from videos captured using monocular cameras. A mixture distribution is computed over the motion orientations and magnitudes of optical flow vectors at each spatial location of the video sequence. Thus, a sequence model which is composed of a direction model and a magnitude model is created by circular and non-circular clustering. Human actions are recognized via a distance metric based on the Bhattacharyya distance that compares the model of a query sequence with the models created from the training sequences. The proposed approach is validated using two public datasets in both indoor and outdoor environments with low and high resolution videos.