Reconnaissance des activités humaines à partir des vecteurs de mouvement quantifiés

Hedi Tabia Michèle Gouiffes Lionel Lacassagne

IEF, Institut d'Electronique Fondamentale
Bat 220, Campus Scientifique d'Orsay
Bures sur Yvette
91405 Orsay, FRANCE

Résumé

Dans cet article, nous proposons une approche pour la reconnaissance des activités humaines à partir des vidéos capturées à l'aide des caméras monoculaires. Nous présentons une méthode basée sur la quantification vectorielle de descripteurs de mouvement. Ces descripteurs sont calculés à partir de l'orientation et la magnitude des vecteurs du flux optique. Nous analysons les performances de deux classifieurs : le classifieur Bayesien Naïf et un classifieur basé sur les Séparateurs à Vaste Marge (SVM). Les résultats montrent l'efficacité de notre approche dans la classification des activités humaines sur la base de données KTH [1].

Mots clefs

Reconnaissance des activités humaines, flux optique, sac-de-mots, Bayesien Naïf, SVM.

1 Introduction

La reconnaissance et la compréhension des actions humaines sont devenues des sujets très populaires dans le domaine de la vision par ordinateur et le domaine du traitement du signal. Un grand nombre d'applications de reconnaissance des actions humaines à partir des vidéos peuvent être trouvées : la vidéo-surveillance, l'interaction homme-machine et l'indexation des vidéos.

Le but d'un système de reconnaissance d'activité humaine est d'identifier les actions simples de la vie quotidienne (comme marcher, courir, sauter ...) à partir des vidéos. Chacune de ces actions, réalisée par une seule personne dans un laps de temps précis, doit être représentée par un modèle de mouvement simple.

Au cours de ces dernières années, de nombreuses méthodes ont été proposées pour la reconnaissance et la compréhension des actions humaines. Elles peuvent être trouvées dans des études bibliographiques complètes tel que [2, 3].

Parmi les travaux de l'état de l'art qui sont directement liés à cet article, citons Laptev et Lindeberg [1], qui ont proposé une méthode de reconnaissance d'actions humaines fondée sur une extraction de caractéristiques spatio-temporelles locales. Ils ont démontré comment des caractéristiques de vitesse adaptées permettent la reconnaissance des actions humaines dans des situations complexes avec mouvements de caméra ou dans des milieux non stationnaires. D'autres caractéristiques spatio-temporelles, comme les *cuboid features* ont été appliquées avec succès dans un système de reconnaissance de comportement humain [4] .

Plus récemment, nous trouvons le travail d'Ali et al. [5] qui proposent d'étudier les actions humaines à partir d'un ensemble de caractéristiques cinématiques issues du flux optique. Ils utilisent une méthode d'apprentissage multi-instance (MIL) pour classifier les actions humaines.

Kosmopoulos et al. [6] ont proposé une approche pour l'étude et la compréhension des comportements visuels en se basant sur l'utilisation de caractéristiques holistiques. Le système proposé a la particularité d'exploiter des informations visuelles provenant de plusieurs caméras.

Plusieurs bases de données sont disponibles pour tester les performances des algorithmes de la reconnaissance des activités humaines. Les plus utilisées sont la bases KTH [1] et la base (AVQ) [7].

Ce papier, propose une méthode de reconnaissance des activités humaines à partir des séquences vidéos enregistrées par une caméra monoculaire. Dans cet article, nous mettons l'accent sur les vidéos provenant de ce type de capteur car il est largement utilisé, s'avère peu moins gourmand en ressources et plus économique.

La méthode présentée ici correspond à un méta-algorithme combinant des histogrammes de flux optique et des classifieurs de sac-de-mots. Les contributions principales de cette approche sont sa simplicité et son efficacité de calcul.

Le reste du papier est organisé comme suit. La section 2 détaille la méthode proposée. Ensuite, dans la section 3 les expériences sont présentées. Nous concluons et proposons plusieurs pistes pour nos travaux futurs dans la Section 4.

2 La méthode

La méthode proposée consiste à représenter une action humaine en fonction d'un ensemble de caractéristiques extraites à partir du flux optique calculé entre deux images successives. Elle comprend quatre étapes principales. 1) La première étape est la construction d'un ensemble d'histogrammes de mouvement pour chaque action humaine présentée dans une séquence vidéo, 2) La deuxième étape consiste à l'arrangement des histogrammes de mouvement en un ensemble fini de clusters prédéterminés (un vocabulaire) à l'aide d'un algorithme de quantification vectorielle, 3) La troisième étape correspond à la construction d'un sac de mots (*keymotions* en anglais), qui tient en compte le nombre d'histogrammes de mouvement affectés à chaque classe. 4) Il faut finalement appliquer un classifieur multi-classes en, considérant le sac de « keymotion » comme un vecteur de caractéristiques, et ainsi déterminer la classe ou les catégories, à laquelle est associée l'action humaine.

Afin d'améliorer la précision de la classification tout en réduisant les temps de calcul, les histogrammes de mouvements construits dans la première étape devraient être suffisamment riches pour discriminer au mieux les différentes classes. Par analogie avec les « keywords » dans le domaine de la catégorisation des documents, nous avons appelé « keymotion » les vecteurs caractéristiques quantifiés qui correspondent aux centres de chaque cluster.

2.1 Extraction des caractéristiques

Préalablement à l'extraction des caractéristiques, un pré-traitement est appliqué sur chaque image de la séquence vidéo. Cette étape est nécessaire pour extraire les objets dynamiques de la scène. La séparation de l'avant plan (les objets en mouvement) et de l'arrière plan est effectuée en appliquant la méthode de mélange de gaussiennes. Á cette fin, nous utilisons une implémentation proposée par [8]. Á cause des résultats inexacts de ce processus, il est nécessaire de raffiner la détection en appliquant une procédure de segmentation. Le but de celle-ci est de supprimer les bruits tout en gardant les objets dynamiques les plus significatifs. Les pixels appartenant au premier plan et ceux du bruit sont séparés par des opérations morphologiques.

Une fois l'ensemble des pixels du premier plan extrait, nous calculons les vecteurs de flux optique.Cette méthode ne nécessite que deux images consécutives pour estimer le mouvement, ainsi des nouvelles caractéristiques peuvent être extraites sur chaque image. L'algorithme de Kanade-Lucas-Tomasi [9, 10] hiérarchique a été choisi parce qu'il représente un bon compromis entre robustesse et temps de calcul. Les vecteurs de mouvement obtenus peuvent s'écrire sous la forme suivante :

$$V_t = \{v_1, ..., v_N | v_i = (\theta_i, D_i)\} \quad (1)$$

où i représente le pixel localisé en (x_i, y_i), θ_i représente l'orientation du mouvement et D_i correspond à la distance entre le pixel i dans l'image à l'instant t et son correspondant dans l'image à l'instant $t + 1$.

Un histogramme de mouvement à deux dimensions est construit pour chaque image de la séquence vidéo. Cet histogramme compte le nombre d'occurrences des vecteurs du flux optique ayant les même valeurs de direction et de magnitude. En effet ces vecteurs de la même image sont arrangés en $N_D \times N_\theta$ « bins », selon la magnitude et la direction, respectivement. La figure 1 montre la structure des histogrammes des flux optiques construits à partir des images extraites de quatre séquences vidéos différentes correspondant à quatre actions humaines différentes.

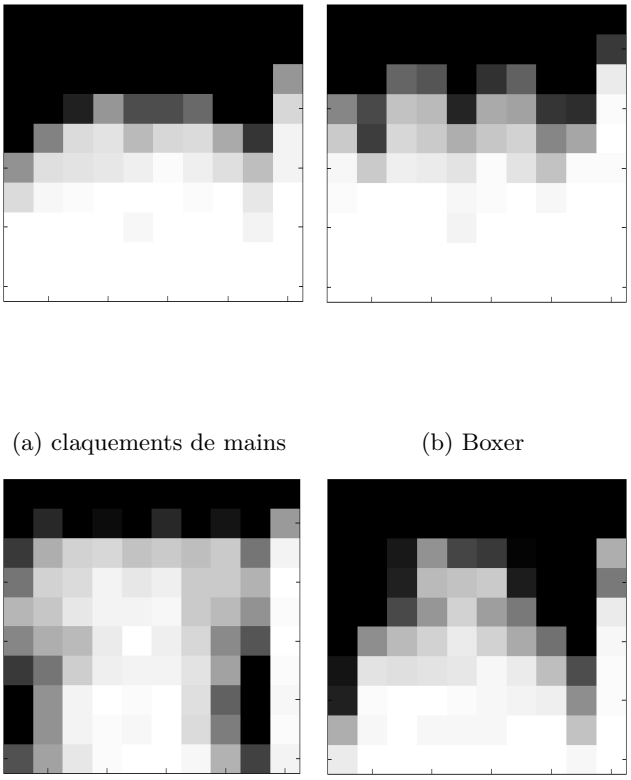
Les lignes des histogrammes de cette figure correspondent à la magnitude du mouvement et les colonnes correspondent à la direction du mouvement. Dans cette visualisation, l'intensité de chaque élément (i, j) est inversement proportionnelle au nombre de vecteurs de flux optique dans le bin (i, j). Les cases noires présentent un important nombre de vecteurs de flux, tandis que les plus clairs indiquent une faible densité de vecteurs mouvement. Afin de simplifier cette représentation, l'histogramme 2D est restructuré dans un histogramme de mouvement à une seule dimension selon cette formule :

$$H(D + \theta \times N_D) = h(D, \theta) \quad (2)$$

La similarité entre deux histogrammes de mouvement est mesurée en utilisant la distance χ^2.

2.2 Vocabulaire de mouvements

Dans l'approche de *sac de mots*, le vocabulaire est obtenu par la quantification de l'ensemble des descripteurs extraits dans la phase d'apprentissage. Le vocabulaire est utilisé pour construire des représentants discriminants, avec lesquels toute action humaine peut être décrite. La méthode la plus utilisée pour construire un vocabulaire d'action est de regrouper les histogrammes rencontrés dans la phase d'apprentissage en un nombre fini de clusters à l'aide d'un algorithme de quantification. Le nombre final de clusters représente la taille du vocabulaire. Pour cette fin, nous avons choisi d'utiliser l'algorithme *k-means*. Il procède par itération en affectant chaque descripteur au *cluster* le plus proche, puis recalcule les centres des *clusters*. L'algorithme *k-means* est appliqué plusieurs fois avec différents nombres de vecteurs représentatifs désiré (k) et différents ensembles de centres de classes initiaux. Nous sélectionnons le (k) final donnant le moins de risque empirique dans la catégorisation.



(a) claquements de mains (b) Boxer

(c) Courir (d) Marcher

Figure 1 – *Histogrammes de mouvement construits pour différentes actions humaines.*

2.3 Reconnaissance d'une action humaine

Après avoir attribué chaque histogramme de mouvement à son plus proche *cluster*, le problème de la reconnaissance des actions humaines peut être ramené à un problème de classification supervisée. Afin de prendre une décision sur une action donnée, le système effectue deux étapes : l'apprentissage et le test. Le but de la phase d'apprentissage est de construire un ensemble de règles qui seront utilisées pour la reconnaissance des actions futures. En effet, en se basant sur les données étiquetées (la vérité terrain), le système est capable de construire des règles de décision pour pouvoir distinguer entre les différents catégories d'actions humaines. En appliquant ces règles de décision sur une action donnée, le système est capable de prédire sa classe. Dans ce papier, nous analysons le comportement des deux classifieurs : Le Bayesien naïf et les Machines à Vecteur de Support.

Classifieur Bayesien Naïf. Le classifier Naïf de Bayes [11] est un classifieur probabiliste basé sur le théorème de Bayes. Pour démontrer le concept de reconnaissance d'action en utilisant ce classifieur, supposons que nous avons un ensemble de séquences étiquetées des actions humaines $S = \{S_i\}$ et et un vocabulaire $A = \{a_t\}$ de *keymotions*. Chaque histogramme de mouvement extrait d'une séquence vidéo est étiqueté avec le *keymotion* auquel il est le plus proche dans l'espace de mouvement. Nous comptons le nombre d'occurrences $N(t, i)$ où le *keymotion* a_t survient dans une séquence vidéo S_i. Pour reconnaître une nouvelle action, la règle de Bayes est appliquée et nous sélectionnons l'activité qui a la plus grande probabilité *a posteriori*.

$$P(C_j/S_i) \propto P(S_i/C_j)P(C_j) = P(C_j)\prod_{t=1}^{|A|} P(a_t/C_j)^{N(t,i)}. \tag{3}$$

Il est évident que dans cette formule que le classifieur naïf de Bayes nécessite une estimation des probabilités conditionnelles de *keymotion* a_t sachant la catégorie d'action donnée C_j. Afin d'éviter les probabilités nulles, ces estimations sont calculées avec un lissage Laplacien :

$$P(a_t/C_j) = \frac{1 + \sum_{S_i \in C_j} N(t,i)}{|A| + \sum_{s=1}^{|A|} \sum_{S_i \in C_j} N(s,i)}. \tag{4}$$

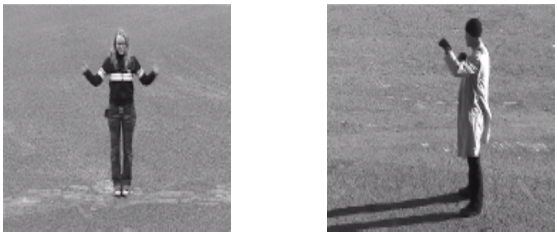
Classifieur SVM. Les machines à vecteurs de support ou séparateurs à vaste marge (SVM) sont un ensemble de techniques d'apprentissage supervisé, destinées à résoudre des problèmes de classification. La classification est réalisées par la construction d'un ensemble d'hyperplans dans un espace multidimensionnel séparant les éléments de différentes classes avec une vaste marge [12]. Afin d'appliquer les SVM multi-classes à notre problème, nous adoptons l'approche *Un-contre-tous*. Étant donné un problème de M classes, nous entraînons m SVM. Chaque SVM est désigné pour discriminer entre une catégorie d'actions donnée i et tous les autres $m - 1$ catégories d'action existantes dans la base.

3 Résultats expérimentaux

Dans cette section, nous présentons les résultats obtenus à partir de deux expériences. Dans la première expérience, nous analysons les performances du classifieur Bayésien naïf et le classifieur basé SVM. Dans la deuxième expérience, nous comparons la performance de notre méthode avec les méthodes de l'état de l'art. Ces expériences ont été réalisées sur un ensemble de données standard contenant une variété d'actions de la vie quotidienne. Les performances de notre méthode ont été évaluées en utilisant 10 validations croisées.

3.1 Description de la base d'actions

Le KTH [1] est une base de données contenant des séquences vidéos de faible résolution (images en niveau de gris avec une résolution de 160 × 120 pixels). La base regroupe six types d'actions humaines effectuées plusieurs fois par 25 personnes. Cette base contient des vidéos en environnement intérieur en extérieur et les personnes portent des tenues vestimentaires différentes.



(a) Claquer les mains (b) Boxer

(c) Courir (d) Marcher

(c) Onduler les mains (d) Jogging

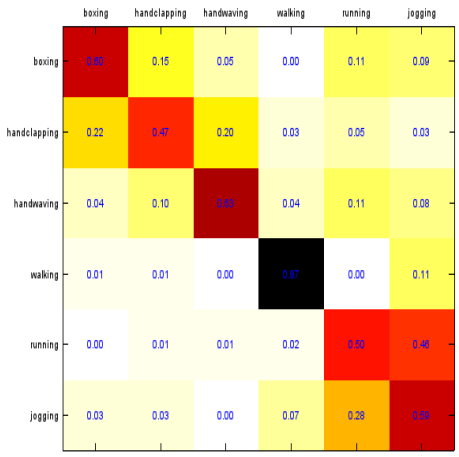
Figure 2 – *Exemple d'actions humaines extraites de la base KTH.*

3.2 Résultat de la classification en appliquant le Bayesien Naïf.

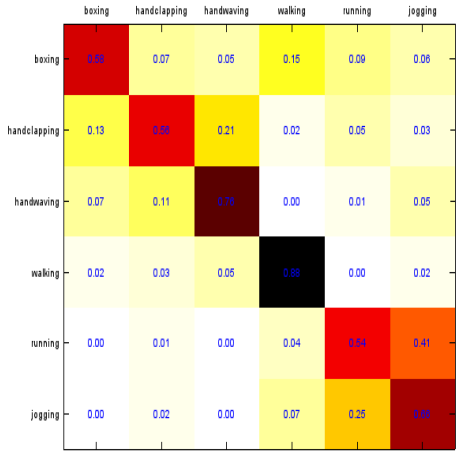
La figure 3 (a) montre les performances de notre méthode en utilisant le classifier Bayesien Naïf. Dans cette visualisation, nous présentons les valeurs correspondantes à la matrice de confusion. Les éléments diagonaux correspondent aux nombres des actions prédites correctement classifiées. Le classifieur Naïf de Bayesien donne 61% comme taux de performance.

3.3 Résultat de la classification en appliquant le SVM

La figure 3 (b) présente les résultats obtenus en appliquant le classifieur SVM. Comme prévu, la performance des SVM surpasse la performance du classifier Naïve de Bayes, en réduisant l'erreur de classification de 39 à 33,67%. Nous avons comparé deux types de classifieur SVM, un premier basé sur un noyau linéaire et un second basé sur un noyau RBF. La méthode basée sur les noyaux RBF offre les meilleures performances.



(a) Avec le classifieur Bayesien Naïf



(b) Avec le classifieur SVM

Figure 3 – *Résultats de la classification des action humaine sur la base KTH.*

En visualisant les deux matrices de confusion de la figure 3, on peut remarquer aussi que les deux classifieurs ont un comportement similaire.

3.4 Comparaison avec les méthodes de l'état de l'art

Afin de mieux évaluer notre approche, nous avons comparé sa performance avec certaines méthodes existantes de l'état de l'art développées pour la reconnaissance des actions humaines. La performance dans cette

Méthode	Taux de performance
Les points d'intérêt espace-temps [13]	0.80
Historiques de vitesse [7]	0.74
Notre méthode (SVM)	**0.66**
Cuboïdes spatio-temporels [4]	0.66
Notre méthode (Naïve Bayesien Naïf)	**0.61**

Tableau 1 – *Comparaison avec l'état de l'art*

section est mesurée en termes de taux de bonne reconnaissance (c'est à dire le pourcentage d'actions qui sont correctement classées). Le Tableau 1 montre que la performance de notre approche est comparable avec les performances d'autres méthodes de l'état de l'art sur la base KHT, bien que les caractéristiques utilisées soient plus simples (histogrammes de vélocité) comparé à ceux de la littérature.

4 Conclusion

Nous avons présenté une méthode de reconnaissance d'actions humaines dans des séquences vidéos. Elle se fonde sur l'analyse de la direction et l'amplitude des vecteurs de flux optique. Une technique de quantification vectorielle est utilisée pour construire des représentants discriminants, avec lesquels toute activité humaine peut être décrite. Afin de prendre une décision concernant la catégorie d'une activité donnée, le système effectue deux étapes : une première étape d'apprentissage et une seconde étape de test. Le but de l'apprentissage est de réaliser un classement correct des activités futures. Basé sur les connaissances apprises sur les données étiquetées, le système est capable d'identifier les catégories des activités dans la phase de test. Dans ce papier, nous avons analysé le comportement des deux classifieurs : le classifieur Bayesien Naïf et le classifieur basé SVM. Les résultats expérimentaux ont montré que la performance des SVM dépasse celle du classificateur Bayesien Naïf. La comparaison avec les méthodes de l'état de l'art a montré que notre méthode est prometteuse, car malgré la simplicité de la représentation de l'action, les résultats obtenus sont comparables avec des méthodes basées sur des représentations spatio-temporelles plus complexes.

Références

[1] I. Laptev et T. Lindeberg. Velocity adaptation of space-time interest points. 2004.

[2] Ronald Poppe. A survey on vision-based human action recognition. *Image Vision Comput.*, 28 :976–990, June 2010.

[3] P. Turaga, R. Chellappa, V. S. Subrahmanian, et O. Udrea. Machine recognition of human activities : A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18, 2008.

[4] P. Dollar, V. Rabaud, G. Cottrell, et S. Belongie. Behavior recognition via sparse spatio-temporal features. Dans *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65 – 72, oct. 2005.

[5] Saad Ali et Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32 :288–303, February 2010.

[6] D. Kosmopoulos et S.P. Chatzis. Robust visual behavior recognition. *Signal Processing Magazine, IEEE*, 27(5) :34 –45, sept. 2010.

[7] R. Messing, C. Pal, et H. Kautz. Activity recognition using the velocity histories of tracked keypoints. Dans *Computer Vision, 2009 IEEE 12th International Conference on*, 2009.

[8] P. Kaewtrakulpong et R. Bowden. An improved adaptive background mixture model for realtime tracking with shadow detection, 2001.

[9] Bruce D. Lucas et Takeo Kanade. An iterative image registration technique with an application to stereo vision. Dans *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, pages 674–679, 1981.

[10] Jianbo Shi et Tomasi. Good features to track. Dans *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 593–600, 1994.

[11] David D. Lewis. Naive (bayes) at forty : The independence assumption in information retrieval. pages 4–15, 1998.

[12] Vladimir N. Vapnik. *Statistical learning theory.* Septembre 1998.

[13] I. Laptev, M. Marszalek, C. Schmid, et B. Rozenfeld. Learning realistic human actions from movies. Dans *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008.