Caractérisation de trajectoires d'activités humaines dans le simplexe sémantique

Cyrille BEAUDRY, Renaud PÉTERI, Laurent MASCARILLA

Laboratoire Mathématiques Image et Applications
Université de La Rochelle Avenue Michel Crépeau, 17042 La Rochelle, France
cyrille.beaudry@univ-lr.fr, renaud.peteri@univ-lr.fr, laurent.mascarilla@univ-lr.fr

Résumé – Cet article présente une approche originale de reconnaissance d'actions humaines complexes dans des vidéos. Elle permet de caractériser au cours du temps des actions élémentaires qui composent une activité (action complexe). Les actions élémentaires sont apprises de façon robuste et générique via une méthode de reconnaissance d'action basée sur l'estimation du flot optique. Elles sont ensuite projetées en tant que trajectoires sur le simplexe des probabilités *a posteriori* des actions élémentaires afin d'être caractérisées et discriminées en tenant compte de la géométrie de ce simplexe. Les résultats obtenus montrent que la méthode permet de différencier différentes classes d'actions humaines.

Abstract – This paper presents an original approach for recognizing human activities in video sequences. This method aims at characterizing human activity as a temporal sequence of elementary actions. These actions are robustly and generically learned using an action recognition method based on optical flow estimation. Finally, activities are projected as trajectories on the simplex of elementary action probability estimated in order to be characterized and discriminated. The metric used in the simplex takes into account its geometry. Results show that the proposed method can differentiate different classes of human activities.

1 Introduction

1.1 Contexte

La reconnaissance d'actions humaines dans des vidéos est une thématique de recherche très active en vision par ordinateur. Elle est utilisée dans beaucoup de domaines, notamment celui de l'indexation automatique de bases de données de vidéos, la vidéo-surveillance, l'intéraction homme-machine, l'analyse de foule, etc. Au niveau sémantique, une activité est un enchaînement temporel d'actions dites élémentaires.
Par exemple, une activité de saut en longueur correspond à l'enchainement des actions : marche, course, saut.

1.2 Reconnaissance d'activités : un bref état de l'art

Dans la littérature, la reconnaissance d'activités humaines est abordée selon deux types de stratégies : par des méthodes discriminatives, largement utilisées dans la reconnaissance d'actions élémentaires, ou par des méthodes génératives probabilistes, qui permettent de mieux prendre en compte la structure temporelle des activités. La plupart des méthodes génératives probabilistes pour la reconnaissance d'activités dans des vidéos s'appuient sur l'Allocation de Dirichlet Latente (LDA) [2]. Cette méthode provient de la recherche de documents et de la fouille de données. 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. Une activité est donc considérée comme une séquence de thèmes au cours du temps. Un classifieur SVM est par la suite souvent utilisé pour discriminer les données obtenues. La découverte de thèmes liées à l'activité humaine a été exploré par Niebles *et al.* [12] comme un apprentissage non supervisé d'actions à partir d'un sac de mots visuels constitué par des descripteurs tels que le descripteur cuboïd [4]. Les séquences vidéos sont subdivisées temporellement, et les thèmes sont découverts sur chaque bloc afin d'y extraire des thèmes représentatifs d'actions élémentaires. Tavernard *et al.*[16] proposent une méthode basée sur la découverte de topics pour la capture de l'information temporelle liée aux activités. Les vidéos sont représentées comme des occurrences, au cours du temps, de mots visuels, élaborés à partir du détecteur spatio-temporel STIP [8]. Un LDA hiérarchique est ensuite utilisé pour prendre en compte la structure chronologique de l'apparition des mots visuels. D'autres approches comme celles de Y.Wang *et al.* ([19] ; [18]) introduisent des versions semi-supervisées du LDA (S-LDA et MM-LDA) afin de forcer la correspondance entre les topics à découvrir et les classes d'actions déjà connues. Ces deux approches sont celles qui présentent les meilleurs taux de reconnaissance(92% sur la base KTH Dataset [15] avec [19] et 83.06% sur UIUC Sport [10] pour [18]). L'inconvénient des processus génératifs tels que LDA, est qu'il n'y a pas de relation directe entre les actions présentes dans les vidéos, qui sont connues, et les topics qui sont découverts de façon non supervisés par l'algorithme LDA. Il est dans ce cadre difficile d'analyser sémantiquement les to-

pics découvert par le LDA. En effet, il n'y a pas de possibilité dans la version originale du LDA, d'apporter une information *a priori* sur les actions déjà connues et d'assurer une correspondance entre les topics générés et les actions présentes dans la vidéo. Les méthodes qui proposent cette information *a priori* ont le défaut d'être bien moins performantes que les méthodes discriminatives classiques sur des bases de données d'actions bien connues de la littérature. De plus, la plupart d'entres elles utilisent des descripteurs globaux, qui ont montrés leurs limites dans le cadre de la reconnaissance d'actions humaines. Dans cet article, nous présentons une approche originale pour la reconnaissance d'activités humaines dans des vidéos. La section 2 présente notre méthode d'extraction de caractéristiques liées aux mouvements dans les vidéos pour construire un classifieur d'actions élémentaires, entrainé sur une base hybride d'apprentissage. La section 3 décrit la transformation des décisions du classifieur en séquence de probabilités d'actions élémentaires. Finalement, la section 4 détaille la projection de ces séquences de probabilités en tant que trajectoires sur un simplexe sémantique d'actions élémentaires afin de caractériser et discriminer les activités en respectant la géométrie de ce simplexe.

2 Reconnaissance d'actions élémentaires

2.1 Une méthode basée sur le flot optique

Nous utilisons ici une méthode issue de nos précédents travaux sur la reconnaissance d'actions humaines [1] basée sur l'utilisation du flot optique. Les vidéos sont caractérisées par des points critiques du flot optique ainsi que par leurs trajectoires au cours du temps. Ces deux éléments sont calculés à différentes échelles spatio-temporelles, suivant une subdivision dyadique des séquences vidéos. Cette subdivision permet d'extraire des points critiques et leurs trajectoires relativement à des mouvements de différentes échelles de fréquences (mouvements rapides et mouvements lents, respectivement, dans les plus hautes et basses échelles de subdivision). Les points critiques estimés sont caractérisés par des descripteurs locaux de formes et de contours (HOG) et d'orientation de mouvement (HOF) [9]. Les trajectoires multi-échelles sont, elles, décrites en utilisant les coefficients de la transformée de Fourier, ce qui garantit une invariance naturelle à différentes transformations géométriques et une description robuste des différentes fréquences de mouvements.Ces trois informations (formes, orientation du mouvement et fréquences) s'avèrent être à la fois complémentaires et pertinentes en terme de taux de reconnaissance. La méthode a prouvé son efficacité sur différentes bases de données d'actions humaines avec des taux de reconnaissance parmi les plus élevés de la littérature [1].

2.2 Apprentissage par jeux de données hybrides

L'objectif ici est de caractériser une activité, ou action complexe comme une succession temporelle des actions qui la composent. Pour cela, Il est nécessaire d'avoir une représentation robuste et générique des actions élémentaires que l'on souhaite extraire. Les bases de données d'apprentissage choisies sont les bases KTH dataset [15], Weizmann dataset [6], UCF-11 [11] et UCF-50 [13]. Elles représentent les actions humaines de différentes façons. KTH et Weizmann, contiennent des vidéos où les actions sont exécutées dans un environnement contraint (caméra statique, fond uniforme, etc). La plupart des ces mouvements sont joués répétitivement et donc de façon peu naturelle (action `boxe` dans KTH, `handwave` dans Weizmann). Ces bases apportent une information canonique concernant les actions élémentaires. Dans UCF-11, les actions sont représentées dans des situations et contextes différents. Ces bases possèdent une plus grande variabilité visuelle et beaucoup plus d'informations non pertinentes (mouvements parasites ne correspondant pas à l'action observée). Ces bases proposent une représentation des actions élémentaires avec une plus grande variabilité. Ces deux catégories de bases de données, avec et sans contraintes, contiennent des informations complémentaires concernant la représentation des actions élémentaires et ce à différentes fréquences de mouvements. Les actions élémentaires considérées dans cet article sont : `Jump`, `Run`, `Walk`, et `Handwave` car elles se retrouvent dans beaucoup d'actions complexes. Entrainer notre classifieur SVM sur ces actions permet d'obtenir une base d'actions élémentaires très peu corrélées. L'étude est effectuée dans un premier temps sur un mélange simple de base de données d'actions. Nous faisons le choix arbitraire d'utiliser une base de données composée au 2/3 de vidéos avec contraintes d'acquisitions (KTH et Weizmann) et 1/3 de vidéos génériques (UCF-11 et UCF-50). Un ensemble de 32 vidéos par classes est ainsi constitué. Le Tableau 1 montre les résultats obtenus après validation croisée en apprenant les actions sur cette base de données. Les seules confusions se font entre classes sémantiquement proches.

TABLE 1 – Matrice de confusion sur le jeu de données hybride

action	$jump$	$walk$	run	$wave$
$jump$	90.62	0	9.37	0
$walk$	0	100	0	0
run	0	3.12	96.87	0
$wave$	0	0	0	100

3 Décomposition d'actions complexes en séquences d'actions élémentaires

3.1 Probabilités d'actions élémentaires

Pour évaluer au cours du temps la proportion des actions élémentaires précédemment sélectionnées et apprises, on utilise les probabilités estimées du SVM, qui sont obtenues par une transformation des décisions du classifieur en probabilités *a posteriori* selon la méthode proposée par [5] et utilisée notamment dans [3]. Ces probabilités sont estimées au cours du temps par fenêtre glissante. L'idée est de signer une frame t de la vidéo en analysant les probabilités des actions élémentaires sur une fenêtre temporelle de taille $[t-N; t+N]$ en considé-

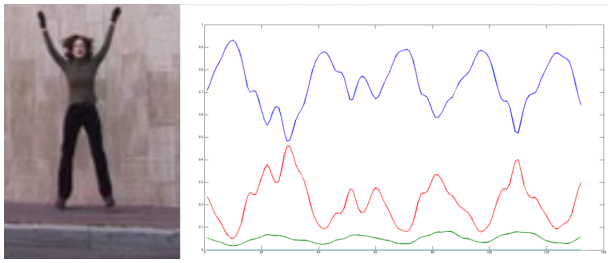

FIGURE 1 – Analyse au cours du temps des probabilités d'actions élémentaires `Handwave`(rouge, `Jump` (bleu), `Walk` (vert) sur une video d'action `Jack` issue de la base Weizmann.

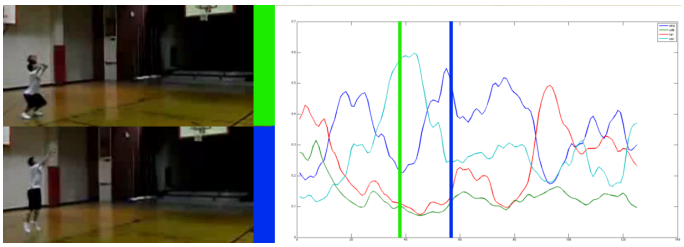

FIGURE 2 – Séquence de basket à deux instants. Au moment où le joueur effectue un mouvement de la main, l'action `Wave` est majoritaire (courbe cyan). Lorsqu'il effectue un tir au panier, l'action `Jump` devient majoritaire (courbe bleue).

rant N petit [14].
La Figure 1 présente l'application de notre méthode sur une vidéo issue de la base Weizmann. L'action exécutée est `Jack`, une action qui est composée à la fois d'un saut alterné avec un mouvement des bras du bas vers le haut. Les courbes du graphe représentent l'évolution de la proportion des actions élémentaires au cours du temps. La courbe rouge correspond à l'action `Handwave`, la courbe bleue à l'action `Jump` et la courbe verte à l'action `Run`. On constate que l'on retrouve à la fois la périodicité du mouvement exécuté sur la séquence et aussi l'alternance entre une proportion forte des actions `Wave` et `Jump`, qui caractérisent à elles deux l'action `Jack`. La Figure 2 illustre l'analyse au cours du temps des probabilités de ces mêmes actions élémentaires sur une séquence vidéo de Basket-ball issue de UCF-11. Ces deux exemples montrent que notre méthode permet une représentation générique des actions apprises lors du mélange de données issues de bases différentes.

3.2 Caractérisation de l'absence d'action

Lors de la phase de reconnaissance des actions élémentaires, la réponse du classifieur dépend de l'action la plus probable parmi celles qui ont été apprises. Il est donc nécessaire d'adapter la réponse lorsqu'il n'y a pas de mouvement présent dans la séquence. La Figure 3 illustre les problèmes que l'on rencontre dans ce cas. Pour évaluer l'absence ou la présence de mouvement, on estime la quantité de mouvement dans la fenêtre temporelle en se basant sur l'énergie du flot optique. Chaque image de la séquence est subdivisée horizontalement et verticalement et l'énergie moyenne du flot optique est calculée sur chacun de ces blocs.
deAu final, chaque image k de la séquence est signée par une valeur $coef_{Standing}(k) \in [\![0, 1]\!]$.

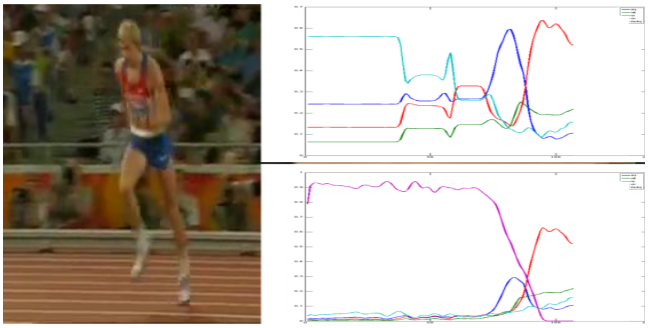

FIGURE 3 – Au début de la séquence, le sportif est immobile. À gauche, sans la classe `Standing`, le classifieur n'arrive pas à interpréter l'absence d'action. À droite, la classe `Standing` est majoritaire et l'action `Run` augmente de façon linéaire au fur et à mesure que le mouvement présent dans la séquence s'intensifie.

Les probabilités d'actions en sortie du classifieur sont renormalisées. En sortie du classifeur, on introduit donc une classe, artificiellement générée, que l'on nomme `Standing`. Cette classe permet de determiner la présence ou non de mouvement dans la séquence à un instant t. Le graphe de la Figure 3 montre un cas avant et après l'utilisation de la classe standing. La création d'une classe artificielle `Standing`, caractérisant l'absence de mouvement permet d'avoir une description plus riche et plus pertinente de la distribution des actions élémentaires au cours du temps dans la séquence.

4 Caractérisation de trajectoires d'actions dans le simplexe

La méthode présentée permet de décomposer une activité en séquences de probabilités d'actions élémentaires. Pour décrire et discriminer ces séquences, nous utilisons le cadre défini pour la reconnaissance de topic sémantique [17]. Nos activités étant définies à partir de séquences de probabilités, les points de ces séquences sont projetés sur le simplexe des probabilités *a posteriori* des actions élémentaires $\mathcal{P}_L$ (L étant le nom d'actions élémentaires) muni de la métrique de Fisher :

$$\mathcal{P}_L = \{\pi \in \mathbb{R}^{L+1} | \sum_{i=1}^{L+1} \pi_i = 1, \pi > 0\},$$

Les activités sont donc représentées comme des trajectoires de points sur ce simplexe sémantique.
Le simplexe $\mathcal{P}_L$ étant muni de la métrique de Fisher, on utilisera l'hypersphère positive $\mathcal{S}_L^+$ de rayon 2, muni de la métrique Euclidienne à sa surface, qui est obtenue par le difféomorphisme $\mathbb{F}$ tel que :

$$\mathbb{F} : \begin{cases} \mathcal{P}_L \mapsto \mathcal{S}_L^+ \\ \pi = (\pi_1, ..., \pi L + 1) \mapsto \theta = (2\sqrt{\pi_1}, ..., 2\sqrt{\pi_{L+1}}) \end{cases}$$

et donc :

$$\mathcal{S}_L^+ = \{\theta \in \mathbb{R}^{L+1} | \sum_{i=1}^{L+1} \theta_i^2 = 2, \theta > 0\},$$

La géodésique entre deux points (π_{k1}, π_{k2}) de $\mathcal{P}_L$ est l'arc de grand cercle reliant $(\mathbb{F}(\pi_{k1}), \mathbb{F}(\pi_{k2}))$ sur $\mathcal{S}_L^+$ [7].
Dans notre expérimentation, nous caractérisons un ensemble de 30 vidéos d'activités sportives, issues des bases de données `UCF-11`, `UCF-50` et `Olympic Sport`. Les trois classes d'acti-

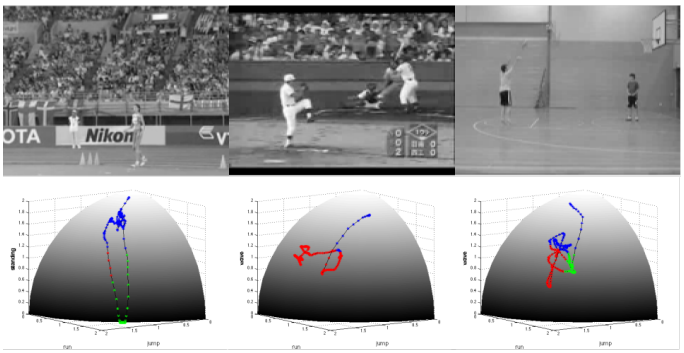

FIGURE 4 – Exemples de trajectoires d'actions de : `High Jump`, `Basket-Ball` et `Basket-Ball`

vités sportives sont : `High-Jump`, `Basket-ball`, `Base-ball`. Ces activités sont caractérisées à l'aide des quatre actions élémentaires `Wave`, `Jump`, `Run` et `Walk`, ce qui fait $L = 5$ en prenant en compte la classe `Standing` (voir Figure 4). Le N de la fenêtre temporelle est fixé à 6.

On constate que ces trajectoires diffèrent entre elles, en terme de position mais aussi en terme de forme. Un autre élément qui tend à discriminer ces trajectoires est l'ordre dans lequel les actions sont enchainées au cours du temps. Ainsi, une trajectoire de `High Jump`, n'est pas temporellement équivalente à `Basket-ball` même si ces deux activités partagent les mêmes actions élémentaires `Wave`, `Jump` et `Run`. Afin d'évaluer la similarité entre les trajectoires des différentes activités, des expérimentations préliminaires ont été réalisées en utilisant la distance de Hausdorff entre trajectoires obtenues sur le simplexe. Cette distance permet d'obtenir un indice de similarité entre deux ensembles fermés de points P et Q par :

$d_H(P,Q) = \max\{\sup_{p\in P} \inf_{q\in Q} d(p,q), \sup_{q\in Q} \inf_{p\in P} d(p,q)\}$

Une classifieur de type *k-plus-proche-voisins*, en utilisant la distance de Hausdorff, a été utilisée sur l'ensemble des données. Le Tableau 2 présente la matrice de confusion obtenue. Le taux global de reconnaissance s'élève à 90%.

Ces travaux préliminaires montrent l'intérêt de caractériser ces activités comme des trajectoires sur le simplexe sémantique à la fois en terme d'analyse visuelle mais aussi en terme de discrimination.

TABLE 2 – Matrice de confusion d'une validation-croisée *Leave-one-out* d'une classification *3-plus-proches-voisins*, suivant la distance de Hausdorff

activités	`High Jump`	`Basket-ball`	`Base-ball`
`High Jump`	90%	0%	10%
`Basket-ball`	0%	100%	0%
`Base-ball`	20%	0%	80%

5 Conclusion

Nous avons ici montré une approche originale de caractérisation d'activités humaines par la décomposition au cours du temps de ces dernières en action élémentaires, représentées comme des trajectoires dans le simplexe sémantique. Les actions élémentaires sont déjà connues et apprises de façon robuste par un classifieur, ce qui est un avantage par rapport aux méthodes génératives probabilistes, qui les découvrent de façon non-supervisée. Les trajectoires d'actions obtenues sont visuellement spécifiques à chaque activité étudiée, ce qui permet de renforcer l'intérêt d'une telle approche. L'utilisation de la distance de Hausdorff est une expérimentation préliminaire, de part le fait que cette distance ne prend en compte ni la géométrie des trajectoires, ni l'aspect temporel des actions. Cependant, les résultats obtenus avec cette distance montre que la représentation actuelle permet d'établir une distinction claire entre des activités étudiées. Dès lors, il est envisageable d'améliorer le processus de discrimination à l'aide d'outils plus spécifiques à la géométrie de la surface utilisée. Les perspectives et travaux en cours portent sur l'élaboration de ces outils, notamment pour la caractérisation de la forme des trajectoires sur l'hyper-sphère positive ainsi que la prise en compte de la structure chronologique des actions élémentaires.

Remerciements Ce travail est partiellement financé par le GDR-ISIS, dans le cadre du projet exploratoire interdisciplinaire TAbaSCO.

Références

[1] Cyrille Beaudry, Renaud Péteri, and Laurent Mascarilla. Action recognition in videos using frequency analysis of critical point trajectories. In *ICIP*, Paris, France, October 2014.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3, March 2003.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011.

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

[5] Ting fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *JMLR*, 5, December 2004.

[6] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *PAMI*, 29(12), December 2007.

[7] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *JMLR*, 6 :129–163, December 2005.

[8] I. Laptev. On space-time interest points. *IJCV*, 64(2-3), 2005.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, June 2008.

[10] Li-Jia Li and Fei-Fei Li. What, where and who ? classifying events by scene and object recognition. In *ICCV*, February 2007.

[11] Jingen Liu, Jiebo Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, June 2009.

[12] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3), September 2008.

[13] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *MVA*, 24(5), 2013.

[14] K. Schindler and L. Van Gool. Action snippets : How many frames does human action recognition require ? In *CVPR*, 2008.

[15] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions : a local svm approach. In *ICPR*, volume 3, 2004.

[16] R. Tavenard, R. Emonet, and J.-M. Odobez. Time-sensitive topic models for action recognition in videos. In *ICIP*, September 2013.

[17] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *PAMI*, 33(11), November 2011.

[18] Yang Wang and Greg Mori. Max-margin latent dirichlet allocation for image classification and annotation. In *BMVC*, April 2011.

[19] Yang Wang, Payam Sabzmeydani, and Greg Mori. Semi-latent dirichlet allocation : A hierarchical model for human action recognition. In *Proceedings of the 2Nd Conference on Human Motion : Understanding, Modeling, Capture and Animation*, pages 240–254, Berlin, Heidelberg, 2007. Springer-Verlag.