Classification des actions humaines basée sur les descripteurs spatio-temporels

Sameh Megrhi, Azeddine Beghdadi, Wided Souidène

L2TI, Institut Galilée, Université Paris 13, Sorbonne Paris Cité,
avenue J. B. Clément 93430 Villetaneuse, France

Résumé. Dans cet article, nous proposons un nouveau descripteur spatio-temporel appelé ST-SURF pour l'analyse et la reconnaissance d'actions dans des flux vidéo. L'idée principale est d'enrichir le descripteur *Speed Up Robust Feature* (SURF) en intégrant l'information de mouvement issue du flot optique. Seuls les points d'intérêts qui ont subi un déplacement sont pris en compte pour générer un dictionnaire de mots visuels (DMV) robuste basé sur l'algorithme des k-moyennes (K-means). Le dictionnaire est utilisé lors du processus d'apprentissage et de reconnaissance d'actions basé sur la méthode des machines à vecteurs supports (SVM). Les résultats obtenus confirment l'intérêt du descripteur proposé ST-SURF pour l'analyse de scènes et en particulier pour la reconnaissance d'actions. La méthode atteind une précision de reconnaissance de l'ordre de 80.7%, équivalente aux performances des descripteurs spatio-temporels de l'état de l'art.

1 Introduction

Le stockage et la distribution à travers les média numériques, et plus particulièrement internet, de données visuelles atteignent des proportions gigantesques. Ceci est accéléré par la banalisation des outils de capture et d'éditions de données numériques telles que la vidéo et l'audio. Cette masse de données de différentes modalités représente une information capitale pour les étapes d'identification d'événements et de décision. Il est donc nécessaire de développer des solutions automatiques pour analyser ces contenus numériques. Une des problématiques qui suscitent beaucoup d'intérêt depuis quelques années est la détection et la reconnaissance des actions humaines dans les séquences vidéo. On appelle action tout évènement caractérisé par des mouvements ou de comportements anormaux que l'on rencontre par exemple dans les flux de vidéo surveillance Bouttefroy et al. (2010). La détection d'actions trouve de nombreuses applications telles que l'indexation, la vidéo surveillance Hu et al. (2004) ou le résumé de vidéos Zhou et al. (2008), pour ne citer que quelques-unes. Dans le contexte de la reconnaissance de l'action, la représentation des descripteurs vidéo au moyen d'un dictionnaire de mots visuels, est un domaine de recherche très actif, Willamowski et al. (2004). L'idée de base d'un DMV est de grouper un ensemble d'objets, par exemple des descripteurs visuels, en groupes de sorte que les objets de même type soient dans un même groupe (cluster). Récemment, l'algorithme des K-moyennes a été largement utilisé pour construire les DMV en

raison de ses hautes performances et de sa simplicité. Chaque vidéo est ensuite représentée par une distribution de mots visuels. Ces distributions servent de paramètres d'entrée dans le processus d'apprentissage à l'issue duquel une classification des actions est obtenue. Cependant, dans une telle approche, la difficulté réside souvent dans la recherche de liens plausibles entre ces entités perceptuelles et l'interprétation de la scène dans le contexte considéré. Il est donc important de trouver le moyen de définir des descripteurs plus riches en information et surtout corrélés aux actions que l'on souhaite identifier et classer.

Afin d'extraire les descripteurs de vidéo, Mojarrad et al. (2008), ont utilisé des descripteurs relatifs à des régions du corps humain, moyennant quelques hypothèses souvent difficiles à satisfaire. Afin, d'éviter ces problèmes, Dollár et al. (2005) ont choisi d'extraire des descripteurs locaux en détectant les cuboïdes locaux, cette méthode produit des mots visuels basés sur la quantification en suivant le même principe que le DMV proposé par Csurka et al. (2004). Dans la même veine, Laptev et Lindeberg (2006) ont proposé le descripteur STIP (Points d'intérêts spatio-temporels) pour détecter les cuboïdes. Néanmoins, les limites des méthodes mentionnées ci-dessus concernent non seulement la difficulté de trouver la taille optimale du "cuboïde", mais aussi le temps de calcul élevé. Pour surmonter ces problèmes, nous proposons un descripteur spatio-temporel basé sur le descripteur local SURF proposé dans Bay et al. (2006). Ce descripteur est ensuite étendu à un espace spatio-temporel 3D. Nous montrons expérimentalement l'efficacité de cette contribution pour la détection des actions humaines dans la base réaliste "UCF sport" proposée par Rodriguez et al. (2008).

2 Approche proposée pour la reconnaissance d'actions

Les points d'intérêt ST-SURF sont localisés à l'aide du détecteur fast-hessien proposé par Beaudet (1978) ensuite extraits à partir de l'intégralité des vidéos de la base d'aprentissage. Les ST-SURF extraits sont regroupés en utilisant l'algorithme de clustering des K-moyennes. Les clips vidéo sont représentés sous forme d'histogrammes de distributions de mots visuels. Enfin, l'étape d'apprentissage est réalisée à l'aide d'une machine à vecteurs de supports (SVM).

2.1 Extention du descripteur SURF dans le domaine temporel

L'extension du descripteur SURF dans le domaine temporel est effectuée en estimant le flot optique proposé par Sun et al. (2010). Ces deniers ont montré que les algorithmes, de calcul du flot optique, fondés sur une étape de filtrage médian permettent d'obtenir un flot optique stable sur un voisinage important, Sun et al. (2010).

Dans cet article, un point d'intérêt $IP = (x, y, t)$ est défini par sa position (x, y) à un instant t . Dans la trame $(t + n)$. Si cet IP effectue un déplacement u suivant la direction x et v suivant celle de y. IP devient, $IP(t + n) = (x + u, y + v, t + n)$. Dans toutes nos expériences, sauf mention du contraire, nous supposons qu'en raison de la segmentation de la vidéo en ensemble de trames (ETr), selon la méthode de Megrhi et al. (2013), les trajectoires des vecteurs de mouvement sont stables et parallèles. Les points d'intérêts tels que $u = v = 0$ seront négligés. L'ensemble des trames forment un volume dans l'espace. Ce volume est un parallélipipède. Ainsi, tout au long des trames du parallélépipède, la direction 3D (u, v, n) représente la direction du mouvement de IP. Le vecteur de mouvement est calculé pour chaque IP. Notre contribution consiste en l'utilisation de l'orientation du mouvement et de sa position

afin de caractériser le mouvement, au lieu d'utiliser le vecteur de direction (u, v, n) généré par le calcul du flot optique. Nous supposons que le vecteur de mouvement dans l'espace 3D peut être défini comme l'intersection de deux plans perpendiculaires respectivement au plan (x, t) et le plan (t, y). Pour extraire l'orientation du IP, nous projetons le vecteur de mouvement sur les plans (x, t) et (t, y) de l'ETr pour définir un angle pour chaque projection, le premier angle $\alpha_x$entre le flot optique et le plan (t, x), l'angle α_y entre le plan (t, y) et le vecteur mouvement.

$$\propto_x = 90 - \frac{180}{\Pi} arctan(u/n), \propto_y = 90 - \frac{180}{\Pi} arctan(v/n). \quad (1)$$

Pour chaque IP, le vecteur de mouvement est projeté sur les plans (t, x) et (t, y) selon deux lignes de supports notées L_x et L_y. La projection orthogonale du centre du parallélipipède, formé par l'ensemble des trames, sur les lignes L_x et L_y permet de calculer deux distances D_x et D_y entre le centre du cube et les lignes supportant les vecteurs de mouvement (L_x et L_y). On obtient alors :

$$D_x = D_{xu} - D_{tv}, D_y = D_{yv} - D_{tu} \quad (2)$$

$$D_x = (x - x_{max}/2)cos(180/\Pi arctan(u/n)) - (t - t_{max}/2)sin(180/\Pi arctan(v/n)) \quad (3)$$

$$D_y = (y - y_{max}/2)cos(180/\Pi arctan(v/n)) - (t - t_{max}/2)sin(180/\Pi arctan(u/n)) \quad (4)$$

avec t_{max}, x_{max} et y_{max} sont les dimensions du Parallélipipède, t_{max} varie en fonction du nombre de trames segmentées. Dans ce qui suit, D_x et D_y décrivent les distances de déplacement d'un point d'intérêt donné. La figure 1, illustre la projection du centre du parallélipipède sur les plans (t, x) et (t, y).

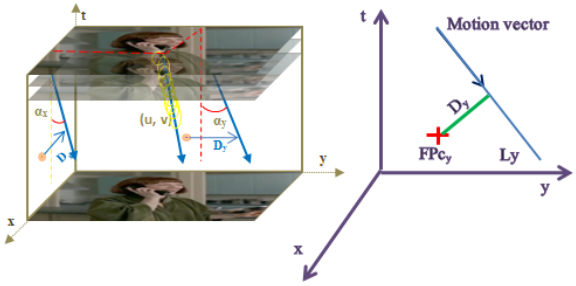

FIG. 1 – *La projection du vecteur mouvement sur les plans adjacents.*

2.2 Extraction du ST-SURF

La génération du nouveau descripteur ST-SURF est réalisée par la fusion par concaténation du descripteur local SURF, de dimension 64-D et les 4 paramètres décrivant la position

et l'orientation de chaque SURF. Le ST-SURF est donc un vecteur de 68-D. Les ST-SURF générés sont quantifiés en mots visuels en utilisant l'algorithme des k-moyennes. Chaque séquence vidéo est alors représentée par l'histogramme de fréquence des mots visuels. Les histogrammes des occurrences de mots visuels qui en résultent sont utilisés comme entrées du classifieur SVM.

3 Résultats expérimentaux

Le tableau 1, comporte les **M**eilleures **M**oyennes de **P**récisions (**MMP**) pour les différents ensembles détecteurs/descripteurs de l'état de l'art ainsi que la **P**récision **M**oyenne reportée en utilisant le détecteur **H**essien (**PMH**). Hu et al. (2004) ont obtenu une précision de 77,4% en utilisant le descripteur HOG et 82,6% en utilisant le HOF. En effet les descripteurs de mouvement local, donnés par les histogrammes des flots optiques(HOF), caractérisent mieux l'action que les histogrammes du gradient orienté (HOG) qui décrivent l'apparence locale. L'utilisation de la combinaison HOG/HOF n'améliore pas la précision de la reconnaissance d'action. En effet, un taux de 81,6% a été reporté par Hu et al. (2004), ceci s'explique par le fait que les HOG sont moins précis pour caractériser l'information temporelle. L'extension du HOG dans le domaine temporel a permis à Wang et al. (2009) d'atteindre 85% en utilisant HOG3D/Gabor. L'orientation spatiale de ce descripteur décrit les informations de l'apparence et l'orientation temporelle extraite renseigne sur la la vitesse du mouvement. La précision du ST-SURF reste en dessous des résultats réalisés par Laptev et al. avec 85 % en utilisant la combinaison HOG3D/Gabor. Ainsi, nous adoptons l'hypothèse que celà pourrait être dû à la génération de différents DMV et l'utilisation de différents détecteurs de points d'intérêt. En utilisant la combinaison détecteur Hessien/SURF, Le ST-SURF que nous proposons donne les meilleurs résultats **PMH** et atteint 80,7 % de précision surpassant tous les descripteurs locaux spatio-temporels de l'état de l'art. En effet, ce descripteur est une combinaison de l'information spatiale, donnée par le SURF, et l'information temporelle dérivée du flot optique.

	HOG3D	**HOG/HOF**	**HOG**	**HOF**	**E-SURF**	**ST-SURF**
MMP	85%	81.6%	77.4%	82.6%	77.3%	**80.7%**
PMH	78.9%	79.3%	66.0%	75.3%	77.3%	**80.7%**

TAB. 1 – *Précision moyenne pour différentes combinaisons de détecteurs/descripteurs appliquées sur la base UCF sport.*

La matrice de confusion relative à la base "UCF sport" est donnée dans le tableau 2. Nous notons que le descripteur proposé donne des résultats satisfaisants dans des vidéos réalistes. Nous soulignons que les précisions les plus faibles sont obtenues par les actions « skate » et « ride », car les mouvements de ces actions sont horizontaux. Le résultat s'améliore au fur et à mesure que les actions contiennent des mouvements verticaux comme « walk », « kick » et « lift » qui présentent des mouvements de rotation importants. Les précisions entre « dive » et « swing » sont trop proches ceci est dû à la ressemblance entre ces deux actions. L'ensemble de nos résultats prouvent que notre méthode est équivalente à l'état de l'art, et montre des performances satisfaisantes sinon meilleures que d'autres méthodes utilisant la même configuration.

TAB. 2 – *Matrice de confusion de la reconnaissance d'actions de la base UCF en utilisant le ST-SURF.*

UCF	Dive	Golf	Kick	Lift	Ride	Run	Skate	Swing	Walk
Dive	0.8	0.17	0	0	0	0	0	0.03	0
Golf	0	0.78	0.2	0	0	0	0	0	0.02
Kick	0	0	0.9	0	0	0.07	0	0	0.03
Lift	0	0	0	0.92	0	0	0	0.08	0
Ride	0	0	0.2	0	0.62	0.18	0	0	0
Run	0	0	0.02	0	0	0.88	0	0	0 .1
Skate	0	0	0.08	0	0	0	0.6	0	0.32
Swing	0	0	0	0	0	0	0.21	0.79	0
Walk	0	0	0	0	0	0.04	0	0	0.96

4 Conclusion

Dans cet article, nous avons proposé un nouveau descripteur spatio-temporel basé sur l'extension du descripteur local SURF vers le domaine temporel. L'extraction du descripteur consiste à détecter des IPs et les projeter dans un espace 3D basé sur une exploitation originale de l'orientation du flot optique et de sa position. Les descripteurs extraits sont intégrés dans un DMV, pour finalement être classés en neuf actions réalistes de la base "UCF sport". En outre, le ST-SURF proposé démontre des performances de reconnaissance prometteuses sur cette base avec une précision d'environ 80,7 %. En effet, le reparamétrage du flot-optique a permis de décrire l'orientation de la trajectoire de la région d'intérêt ainsi que sa position dans un volume spatio-temporel. L'exploitation de l'information relative à l'orientaion garantit l'invariance par rotation du ST-SURF. La position de la région d'intérêt est extraite afin d'améliorer et optimiser la classification pour plus de précision. Ainsi, la classification sera plus robuste aux décalages de pixels qui peuvent aboutir à plus de mots visuels. Ainsi, en utilisant la position de la région d'intérêt à partir du flot optique (au lieu des coordonnées du point d'intérêts) nous obtenons un DMV plus compact. Enfin, les résultats obtenus démontrent la viabilité de notre approche et prouve que nous sommes déjà équivalents aux performances données par l'état de l'art. Nous imaginons de nombreuses perspectives pour l'avenir, la plus importante est d'appliquer la même méthode sur des vidéos contenant des actions plus complexes. Nous prévoyons également d'améliorer notre ST-SURF et envisageons de le combiner avec d'autres descripteurs de bas niveau et de différentes modalités.

Références

Bay, H., T. Tuytelaars, et L. Van Gool (2006). Surf : Speeded up robust features. In *Computer Vision–ECCV*, pp. 404–417. Springer.

Beaudet, P. R. (1978). Rotationally invariant image operators. In *Proceedings of the International Joint Conference on Pattern Recognition*, pp. 579–583.

Bouttefroy, P., A. Beghdadi, A. Bouzerdoum, et S. Phung (2010). Markov random fields for abnormal behavior detection on highways. In *Visual Information Processing (EUVIP), 2010 2nd European Workshop on*, pp. 149–154. IEEE.

Csurka, G., C. Dance, L. Fan, J. Willamowski, et C. Bray (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, pp. 22.

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

Hu, W., T. Tan, L. Wang, et S. Maybank (2004). A survey on visual surveillance of object motion and behaviors. *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, 334–352.

Laptev, I. et T. Lindeberg (2006). Local descriptors for spatio-temporal recognition. In *Spatial Coherence for Visual Motion Analysis*, pp. 91–103. Springer.

Megrhi, S., W. Souidène, et A. Beghdadi (2013). Spatio-temporal surf for human action recognition. In *The Pacific-Rim Conference on Multimedia (PCM)*. LNCS.

Mojarrad, M., M. A. Dezfouli, et A. M. Rahmani (2008). Feature extraction of human body composition in images by segmentation method. *World Academy of Science, Engineering and Technology*.

Rodriguez, M. D., J. Ahmed, et M. Shah (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE.

Sun, D., S. Roth, et M. J. Black (2010). Secrets of optical flow estimation and their principles. In *IEEE Conference onComputer Vision and Pattern Recognition CVPR.*, pp. 2432–2439. IEEE.

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

Willamowski, J., D. Arregui, G. Csurka, C. R. Dance, et L. Fan (2004). Categorizing nine visual classes using local appearance descriptors. *illumination*, 21.

Zhou, Z., X. Chen, Y.-C. Chung, Z. He, T. X. Han, et J. M. Keller (2008). Activity analysis, summarization, and visualization for indoor human activity monitoring. *Circuits and Systems for Video Technology, IEEE Transactions on 18*(11), 1489–1498.

Summary

The goal of this paper is to introduce a spatio-temporal descriptor that we call ST-SURF in order to efficiently describe video data in the context of human action recognition. ST-SURF is the fusion of the *Speed Up Robust Feature* (SURF) and the an original parameterization of the optical flow. Video sequences are represented by a "Visual-bag-of-words" (BoVW). The former (BoVW) is an a video representation which aims to train a support vector Machine (SVM) for action recognition detection. We observe that the ST-SURF leads to relevant action recognition rates on "UCF sport" dataset.