Analyse de trajectoires pour l'indexation sémantique des vidéos à grande échelle

Sabin Tiberius Strat, Alexandre Benoit, Patrick Lambert

Analyse de trajectoires pour l'indexation sémantique des vidéos à grande échelle

Sabin Tiberius STRAT[1,2] Alexandre BENOIT[1] Patrick LAMBERT[1]

[1] LISTIC, Université de Savoie
[2] LAPI, Université "Politehnica" de Bucarest

Sabin-Tiberius.Strat@univ-savoie.fr

Résumé

L'indexation sémantique automatique de grandes collections vidéo est un problème complexe qui ne peut se limiter à l'analyse de mots clefs saisis par les utilisateurs. L'extraction de descripteurs spatiaux et temporels à partir du contenu est devenu indispensable pour appréhender la grande richesse des contenus. Cependant, le coût de calcul associé à l'extraction de descriptions temporelles est bloquant et les systèmes actuels se limitent souvent au traitement de l'information spatiale. Dans ces travaux, nous proposons une méthode d'analyse de trajectoires de points caractéristiques. Ses besoins en ressources de calcul sont faibles et s'adaptent facilement à de grandes collections vidéo. Partant d'une description des trajectoires de points d'intérêt utilisant l'analyse du flot optique, une batterie de descripteurs basés sur les modèles Sac de Mots sont calculés pour représenter les vidéos. L'approche est évaluée dans le contexte difficile du challenge TRECVid Semantic INdexing (SIN).

Mots Clef

trajectoires, suivi, vidéo, indexation sémantique, Sac de Mots

Abstract

A generic system for semantic indexing of large video datasets requires the extraction of both spatial and temporal descriptors from the video content. However, the computational cost associated with the extraction of temporal descriptions causes most approaches to limit themselves to a spatial description. We propose a motion description approach based on trajectories of tracked points that keeps computational cost low, therefore it can be better scaled to large datasets. Our approach constructs trajectories of interest points by tracking them across many frames using optical flow, computing a battery of motion descriptors along each trajectory and constructing Bag of Words representations. We evaluate our approach in the difficult context of the TRECVid Semantic INdexing (SIN) challenge.

Keywords

trajectories, tracking, video, semantic indexing, Bag of Words

1 Introduction

Les dernières années, nous nous sommes confrontés avec une forte augmentation de la quantité de données vidéo, ce qui demande des outils automatiques d'indexation sémantique par le contenu. Ces outils impliquent une étape d'*extraction de descripteurs*, qui permet d'extraire à partir des vidéos des représentations compactes et compréhensibles par l'ordinateur. Elle est suivie d'une étape de *classification supervisée* permettant de reconnaître des concepts sémantiques de haut niveau à partir de ces représentations. Dans une phase d'optimisation, des techniques de *fusion d'informations* sont ajoutées lorsque l'on dispose de descripteurs complémentaires.

Une famille de descripteurs performante et couramment utilisée est celle des *Sacs de Mots Visuels* (*Bag of Visual Words, BoW*) [1], qui représente les objets ou d'autres concepts comme une collection non-ordonnée de motifs redondants (les "mots visuels"). Initialement développée pour l'analyse des images statiques, cette méthode a été adaptée aux vidéos. Cependant, les performances de cette approche dans des campagnes d'évaluation de systèmes d'indexation vidéo plus complexes, comme MediaEval [2] et TRECVid [3] restent assez faibles. Ceci est lié à la variabilité supplémentaire apportée par la dimension temporelle qui s'ajoute à la grande diversité sémantique des données de ces campagnes. Par ailleurs, le coût de calcul supplémentaire demandé pour l'analyse temporelle des vidéos, est souvent bloquant. En conséquence, la plupart des approches décrivent uniquement quelques images-clef [4] ignorant ainsi complètement l'information temporelle.

Ce travail se focalise donc sur l'inclusion de l'information temporelle dans la description des vidéos, en complément des méthodes spatiales, pour obtenir un système d'indexation plus générique, capable d'appréhender des concepts liés au mouvement. Notre méthode est basée sur le modèle Sac de Mots (BoW), mais nous remplaçons les caractéristiques locales classiquement *spatiales* par des trajectoires de points d'intérêt. Ces trajectoires sont regroupées par un clustering de type "Kmeans" dans des "mots-trajectoire", et des histogrammes BoW sont utilisées pour représenter chaque élément vidéo.

Nous testons notre approche sur la collection d'entraî-

nement du challenge TRECVid 2012 Semantic Indexing (SIN). Cette collection comprend 346 concepts sémantiques divers, comme des objets (Carte, Voiture, Chien), des types de scène (Nuit, Intérieur, Paysage urbain), des actions (Danser, Courir, Boire), des personnes (Bill Clinton, Personnes asiatiques) etc. dans environ 200 heures de vidéo acquises dans différentes conditions (contexte, éclairage, mouvement de la caméra) et avec différents niveaux de qualité d'encodage. Les vidéos sont pré-découpées en environ 400 000 plans de courte durée dans lesquels les concepts sémantiques doivent être détectés. Le volume et la variabilité des collections TRECVid SIN est un contexte d'expérimentation difficile mais a l'avantage d'être très réaliste.

Le reste du papier est structuré de la manière suivante : La section 2 présente l'état de l'art pour la description du mouvement, la section 3 décrit notre approche, la section 4 donne nos résultats expérimentaux et la section 5 conclut le papier.

2 Etat de l'art

Différentes méthodes de description de l'information temporelle des vidéos sont possibles. On peut distinguer 2 catégories : les méthodes globales et locales. Les méthodes globales décrivent des volumes vidéo dans leur intégralité, tandis que les approches locales décrivent de plus courts éléments et sont plus robustes aux occultations. Pour cette raison, les approches locales sont préférées dans des contextes non contrôlés comme TRECVid SIN. Une stratégie BoW est généralement utilisée pour agréger des caractéristiques locales pour chaque plan vidéo. Les méthodes reconnus de l'état de l'art sont les points d'intérêt spatio-temporels [5], MoSIFT [6], ou les *trajectoires de points d'intérêt*. Nous sommes intéressés par cette dernière catégorie, car les trajectoires peuvent décrire des mouvements plus longs que les autres approches. Néanmoins, les trajectoires restent des approches locales, car leur longueur est courte comparée à la durée d'un plan vidéo.

Parmi les méthodes utilisant les trajectoires, certaines utilisent un suivi basé sur le flot optique dense [7, 8, 9], tandis que d'autres suivent un jeu limité de points d'intérêt ce qui réduit le coût de calcul [10, 11]. Une trajectoire peut ensuite être décrite par des vecteurs de déplacement [10], d'accélération [11], ou par des Histogrammes d'Orientation du Gradient (HOG), du Flot Optique (HOF) ou des Frontières du Mouvement (MBH, Motion Boundary Histograms) qui décrivent le voisinage de la trajectoire [8, 9]. Les MBH sont robustes au mouvement de la caméra [12]. Cependant, la précision de la compensation du mouvement de la caméra est un point critique et difficile à gérer pour des vidéos non-contrôlées. Les méthodes de l'état de l'art sont décrites dans [7, 8, 9].

Une fois les trajectoires décrites, un modèle d'agrégation comme les sacs de mots est souvent utilisé. Les travaux comme [10, 11] ne considèrent pas les relations spatiales et/ou temporelles entre les trajectoires. Une autre approche, l'Actom Sequence Model [13] prend en compte la succession temporelle des éléments d'action et donne de meilleurs résultats. Cependant, des annotations plus détaillées sont requises en amont, dans la phase d'apprentissage.

D'un point de vue général, très peu de ces contributions sont testées sur des collections génériques comme TRECVid SIN, la plupart se limitant aux collections spécialisées sur les actions et dans des contextes plus contrôlés (films, sports olympiques télévisés etc.) et richement annotés. A notre connaissance, les seuls travaux auxquels nous pouvons comparer notre approche sur TRECVid SIN sont ceux de [4], où un descripteur de mouvement basé sur des trajectoires de points d'intérêt dans un modèle Sac de Mots est proposé. De plus, le même classifieur supervisé est utilisé que dans notre approche, ce qui facilite la comparaison.

3 Approche proposée

Notre approche combine une série d'améliorations de l'état de l'art sur les trajectoires de points d'intérêt. Elle ajoute des modifications qui permettent l'indexation de très grandes collections ayant un contenu riche, comme TRECVid SIN. Nous proposons une approche équilibrée en terme de coût de calcul, richesse de la description du mouvement et robustesse aux mouvements de la caméra.

3.1 Construction des trajectoires

Pour construire des trajectoires, nous détectons des points d'intérêt "Good Features To Track" (GFTT) [14]. Nous suivons ces points le long de la séquence vidéo en utilisant le flot optique calculé à l'aide de l'algorithme de Lucas-Kanade [15]. Même si un échantillonnage dense des points peut donner de meilleurs taux de précision dans les approches Sac de Mots, nous avons opté, pour des raisons liés à l'implémentation, pour des points GFTT. Des travaux futurs adresseront l'échantillonnage dense. Nous limitons la longueur maximale d'une trajectoire à 2 secondes (50 images pour les vidéos TRECVid), durée pour laquelle le cumul des erreurs de suivi reste à un niveau acceptable. Un exemple de trajectoires obtenues est donné à la Figure 1.

Tout au long du calcul des trajectoires, nous estimons le mouvement de la caméra en calculant le flot optique pour une grille régulière de 1000 points. L'homographie correspondante est estimée en utilisant l'algorithme RANSAC [16] pour exclure les valeurs aberrantes. Notre méthode est une version simplifiée de [8], afin de maintenir un coût de calcul plus faible.

A la fin de la séquence vidéo, pour chaque trajectoire, les extrémités sont éliminées si elles sont statiques (si entre deux images successives, le déplacement est inférieur à 10% du déplacement maximal au long de la trajectoire). Ce traitement améliore les résultats d'environ 5-20%, en fonction de la description utilisée pour les trajectoires.

Notre stratégie permet le traitement en temps réel : une vidéo de 320x240 pixels à 25 images/s, d'une durée de 15s, est traitée en 14s, donnant 35000 trajectoires. Le test a été

FIGURE 1 – Exemples de trajectoires pour un skateboarder en train de tourner : les trajectoires le suivent lors de son mouvement vers la gauche et du changement de direction vers la droite

effectué sur un PC doté d'un processeur Inter Core i5 M560 à 2,67GHz.

3.2 Description des trajectoires

Chaque trajectoire est décrite avec une batterie de descripteurs :

1. *histogramme des directions du mouvement* des points d'intérêt, avec 8 bins pour l'orientation et pondération de chaque mouvement entre deux images successives par son amplitude ;
2. histogramme des directions du mouvement des points d'intérêt avec un *bin pour le mouvement nul*, sans pondération ; le seuil pour le mouvement nul est fixé à 20% du mouvement maximal le long de la trajectoire ;
3. *histogramme des directions d'accélération*, similaire à (1) pour les accélérations ;
4. histogramme des directions d'accélération avec un bin pour *l'accélération nulle*, similaire à (2) pour les accélérations ;
5. vecteur de 8 *vitesses (composantes x,y)* le long de la trajectoire ; obtenu par ré-échantillonnage de la trajectoire et normalisation par rapport au mouvement total de la trajectoire ;
6. similaire à (5) mais avec 16 échantillons de vitesse ;
7. vecteur de 7 *accélérations (composantes x,y)*, obtenu par dérivation de (5) avec normalisation par rapport à l'accélération totale de la trajectoire ;
8. similaire à (7) mais avec 15 échantillons d'accélération ;

Toutes les représentations proposées sont robustes à l'échelle temporelle et spatiale. Cependant, la durée d'un mouvement est une information pertinente. Par conséquent, nous concaténons à chaque représentation ci-dessus la longueur de la trajectoire (exprimée en secondes). Des expérimentations préliminaires ont montré que cette information apporte une amélioration des résultats d'environ 3 à 4% (jusqu'à 16% pour l'histogramme d'orientations de l'accélération avec un bin zéro).

3.3 Agrégation dans le modèle Sac de Mots

Dans TRECVid SIN, on ne dispose d'annotations qu'au niveau des plans vidéo. Par conséquent, nous utilisons un modèle générique de type *Sac de Mots (BoW)* [1] pour regrouper les trajectoires, ce modèle ignorant les relations entre les trajectoires. Nous ne pouvons pas appliquer un modèle plus complexe comme l'Actom Sequence Model [13] car nous ne disposons pas d'annotations précises sur la séquence des éléments d'actions pour la phase d'apprentissage. Dans l'approche BoW que nous utilisons, chaque type de description de trajectoire donne lieu à un modèle BoW différent et est traitée indépendamment des autres descriptions.

Pour chaque type de description de trajectoire, un vocabulaire est appris en utilisant le clustering "Kmeans", avec la méthode d'initialisation "Kmeans++" [17], appliqué à un ensemble de trajectoires issues de la collection d'entraînement. Ensuite, les plans vidéo sont décrits comme des histogrammes de *"mots-trajectoire"*, comme dans les approches BoW classiques. Nous associons directement une trajectoire à son mot-trajectoire le plus proche. Plusieurs tailles de vocabulaire ont été testées pour chaque description, mais nous ne présentons que les meilleurs résultats dans le Tableau 1.

4 Expérimentations

Nous évaluons notre approche sur la collection de développement de TRECVid SIN 2012. Pour chacun des 346 concepts cibles, le but est de retourner une liste ordonnée d'un maximum de 2 000 plans vidéo susceptibles de contenir le concept. La liste est évaluée en utilisant la mesure officielle de TRECVid SIN, la précision moyenne par inférence (infAP) [18].

La collection est divisée en deux parties, *2012x* et *2012y*. Chacune contient environ 100 heures de vidéo prédécoupées en environ 200 000 plans. Les vocabulaires de mots-trajectoire et l'entraînement de classifieurs supervisés se font sur *2012x*, l'évaluation des performances se fait sur *2012y*. Pour la classification supervisée, nous utilisons un classifieur KNN (K plus proches voisins) de [4], très rapide car il permet de trouver les N voisins une seule fois pour l'ensemble des 346 concepts. Un classifieur plus complexe comme MSVM (SVM avec apprentissage multiple) de [4] donne des meilleurs résultats (un gain d'environ 10-15% peut être attendu dans ce contexte [4]), mais il a un coût de calcul nettement plus élevé.

4.1 Résultats globaux

Le Tableau 1 montre la performance moyenne (infAP) sur les 346 concepts de TRECVid SIN. Parmi ces 346 concepts, moins de 100 semblent, intuitivement, être liés

TABLE 1 – Résultats globaux sur TRECVid SIN 2012y, exprimés en précisions moyennes par inférence (infAP), en moyenne sur tous les 346 concepts. K est la taille du vocabulaire du Sac de Mots. "c.c." représente la compensation du mouvement de la caméra.

Descripteur trajectoire	**K**	**AP**	**AP c.c.**
hist. dir. mouv.(1)	256	0.0391	0.0386
hist. dir. mouv. bin-0 (2)	256	0.0367	0.0282
hist. dir. accel.(3)	256	0.0408	0.0386
hist. dir. accel. bin-0 (4)	256	0.0311	0.0254
vect. vitesses 8 éch. (5)	384	0.0385	**0.0425**
vect. vitesses 16 éch. (6)	384	0.0386	**0.0419**
vect. accél. 7 éch. (7)	384	0.0413	0.0412
vect. accél. 15 éch. (8)	768	**0.0444**	0.0436
Pour comparaison :	**K**	**AP**	**-**
descr. mouv. de [4]	1000	0.0528	-
SIFT retina statique [19]	1024	0.0989	-
Fusions tardives de :	**-**	**AP**	**-**
(1)-(8) basique + c.c.	-	**0.0805**	-
famille SIFT retina [19]	-	0.1366	-
fusion des 2 précédents	-	**0.1445**	-

TABLE 2 – Résultats concept par concept sur TRECVid 2012y, exprimés en infAP. Pour comparaison, nous incluons le descripteur statique *SIFT retina* de [19] et la performance d'un classifieur aléatoire. Les noms originaux des concepts ont été conservés pour éviter toute confusion.

Concept	(1)	(5) c.c.	SIFT r.	aléat.
Eaters	**0.1044**	0.0633	0.0428	0.0028
Indoor sports v.	0.0925	**0.2103**	0.1802	0.0163
Fight - physical	0.0087	**0.0661**	0.0194	0.0046
Football	**0.0982**	0.0715	0.0519	0.0021
Pickup truck	**0.1340**	0.1255	0.1286	0.0019
Rifles	0.0621	**0.0952**	0.0500	0.0107
Court	**0.0588**	**0.0588**	0.0064	0.0004
Press conf.	0.0127	**0.0851**	0.0144	0.0109
First lady	0.1409	0.1369	**0.1559**	0.0008
Chair	0.1052	0.1285	**0.1468**	0.0460
Female reporter	0.0581	0.1357	**0.1976**	0.0076
Van	0.0928	0.0713	**0.1391**	0.0026
Running	0.1224	0.1257	**0.1509**	0.0064
Soccer player	0.2310	0.2303	**0.3096**	0.0020
Throwing	0.1195	0.1276	**0.1984**	0.0037
Skating	0.0384	0.1133	**0.1525**	0.0240

au mouvement. Par conséquent, les résultats ne peuvent dépasser ceux obtenus avec une description purement spatiale. Toutefois, nos résultats sont du même ordre de grandeur que ceux obtenus par d'autres approches. Par exemple, sur la même collection et en utilisant le même classifieur KNN, un descripteur assez performant basé sur un Sac de Mots de caractéristiques locales spatiales SIFT, utilisant un modèle de rétine pour pré-traiter les vidéos (*SIFT retina*) obtient 0.0989 infAP [19]. Cet ordre de grandeur est normal pour cette collection difficile et pour laquelle un seul descripteur ne peut pas décrire tout l'ensemble de concepts sémantiques. Par rapport à d'autres approches basées sur le mouvement, les résultats de notre meilleur descripteur (0.0444) sont comparables avec ceux obtenus par le descripteur de mouvement dans [4] (0.0528 infAP), qui utilise les mêmes collections *2012x* et *2012y* pour l'entraînement et le test.

Globalement, le meilleur descripteur est le vecteur d'accélérations avec 15 échantillons (8), sans compensation du mouvement de la caméra. La compensation du mouvement de la caméra améliore les résultats basés sur les vecteurs de vitesse, mais en général, les résultats sont inférieurs. Le mouvement de la caméra ne peut pas être toujours déterminé correctement, surtout quand les objets véritablement en mouvement occupent une grande partie de l'image.

4.2 Résultats concept par concept

Même si un certain descripteur peut être le meilleur globalement, ce n'est pas toujours le cas pour chaque concept. Le Tableau 2 montre cette situation, en comparant les meilleurs descripteurs de trajectoire avec le descripteur spatial *SIFT retina* de [19]. Le tableau est structuré en 4 groupes : le premier contient des concepts qui sont mieux détectés avec les descripteurs de trajectoire grâce à leur liaison avec le mouvement ; le deuxième montre quelques résultats inattendus, avec des concepts plutôt spatiaux, mieux détectés avec les trajectoires ; le troisième montre des concepts spatiaux, assez bien détectés avec les trajectoires ; le dernier montre des concepts liés au mouvement qui sont mieux détectés avec le descripteur statique.

En continuant notre analyse, pour les descripteurs de trajectoire du Tableau 2, 142 concepts sur les 346 ont été mieux détectés qu'un classifieur aléatoire. Ce résultat est encourageant, car moins de 100 concepts semblent liés au mouvement. Parmi ces 142 concepts, 29 sont mieux détectés avec l'un des descripteurs de trajectoire qu'avec le descripteur statique. Cela montre que les trajectoires apportent de l'information même dans le cas difficile de TRECVid (contexte très divers, mouvement de la caméra, duplication des images au re-encodage des vidéos etc.).

Une conclusion préliminaire est qu'il est préférable d'avoir une grande batterie de descripteurs. En effet, la plus grande partie du coût de calcul provient du calcul des trajectoires, les descripteurs de trajectoire étant très simples à calculer. Comme aucun descripteur ne peut être le meilleur pour tous les concepts, la question de la fusion des descripteurs se pose en vue d'améliorer les résultats.

4.3 Fusion tardive de scores de classification

Nous réalisons des fusions tardives des scores de classification (pour chaque plan vidéo et chaque concept) obtenus avec les différents descripteurs de trajectoire (1)-(8). L'algorithme de fusion, décrit dans [20], basé sur la méthode

AdaBoost, donne de bons résultats dans ce contexte. Les résultats sont présentés à la fin du Tableau 1.

La fusion tardive des descripteurs (1)-(8) double quasiment les performances moyennes par rapport au meilleur descripteur individuel, ce qui montre que nos descripteurs sont vraiment complémentaires. Comparé au descripteur de mouvement de [4], nos performances sont 52% plus grandes. Cela montre que notre batterie de descripteurs simples est meilleure qu'un seul descripteur spécialisé. Ceci est d'autant plus intéressant que le coût de calcul de la fusion tardive est négligeable par rapport au coût d'extraction des descripteurs. Dans nos expérimentations sur la base TRECVid SIN, nous avons trouvé qu'une fusion précoce des descripteurs est moins bonne que la fusion tardive, permettant d'atteindre seulement 0.0483 infAP,

Évidemment, une fusion de descripteurs plutôt spatiaux (la famille SIFT dans [19]) reste meilleure qu'une fusion de descripteurs de trajectoires car la plupart des concepts ne sont pas liés au mouvement. Toutefois, un niveau supplémentaire de fusion, combinant par une simple moyenne arithmétique, les scores de classification de la fusion de la famille spatiale de [19] et les résultats de la fusion des descripteurs de trajectoires, améliore encore de 5.8% les performances globales. Cela montre que la complémentarité entre les descripteurs spatiaux et de mouvement peut être exploitée pour améliorer l'indexation sémantique d'un ensemble divers de concepts.

4.4 Expérimentations sur la base KTH

Afin de valider la généricité de notre approche, nous avons réalisé une expérimentation sur la base KTH [21]. Cette base est dédiée à la reconnaissance d'actions. Elle en comporte 6 (boxer, applaudir, agiter les mains, trotter, courir, marcher) réalisées par 25 personnes différentes dans des contextes simples.

Nous avons utilisé les vocabulaires de mots-trajectoires extraits sur la base TRECVid 2012x pour représenter les vidéos de la collection KTH avec l'approche Sacs de Mots. Pour l'entraînement du classifieur final, nous avons utilisé les vidéos des 8 premières personnes, et nous avons utilisé les vidéos des 9 dernières pour le test. La procédure est ainsi similaire à [21]. Cependant, nous n'avons pas utilisé les 8 personnes restantes pour faire des validations croisées, afin de rester dans le cadre d'un système d'indexation générique non spécialisé pour une collection particulière.

Le classifieur supervisé utilisé est le 3-ppv (3 plus proches voisins). Un classifieur plus complexe peut améliorer les résultats, mais nous avons opté pour un classifieur de la même famille que celui utilisé pour TRECVid SIN, afin de garder des conditions de test similaires.

Le Tableau 3 montre les résultats obtenus pour deux descripteurs : le premier consiste à décrire une trajectoire par la concaténation des descripteurs (1,2,3,4) du Tableau 1, et le deuxième par la concaténation des descripteurs (5,6,7,8). Même si les résultats sont inférieurs aux dernières contributions trouvées dans l'état de l'art, nous rappelons que nous n'avons réalisé aucune optimisation de paramètres pour la base KTH, notre but étant de tester la généricité de l'approche. Toutefois, nous pouvons remarquer que les vocabulaires extraits sur TRECVid donnent des performances assez proches de celles obtenues avec des vocabulaires extraits spécifiquement sur KTH. En conclusion, le vocabulaire issu des donnés génériques TRECVid, même s'il provient d'une collection totalement différente, a du sens et peut être utilisé directement sur d'autres bases vidéo.

TABLE 3 – Précisions de reconnaissance d'actions sur la base KTH, avec des vocabulaires de mots-trajectoire extraites sur TRECVid SIN et sur KTH. K est la taille du vocabulaire du Sac de Mots. TV et KTH représentent les vocabulaires extraites sur TRECVid et KTH respectivement.

Descripteur de trajectoire	**K**	**v. TV**	**v. KTH**
concaténation (1,2,3,4)	64	77%	72%
concaténation (5,6,7,8)	192	71%	77%
Pour comparaison :	**-**	**-**	**P(%)**
trajectoires denses [12]	-	-	94.2%
points spatio-temporels [21]	-	-	71.7%

5 Conclusion

Ce papier montre que la description des trajectoires de points d'intérêt est une méthode viable pour la détection de concepts sémantiques divers dans des scénarios vidéo non contrôlés, en complément aux descripteurs spatiaux. Grâce à la stratégie de suivi et à la simplicité de nos descripteurs, le coût de calcul reste assez faible et l'approche est adaptable aux très grandes collections. L'approche est très générique : au lieu de construire un seul descripteur optimisé, une batterie de descripteurs simples est construite et une fusion tardive de très faible coût de calcul permet de dépasser l'état de l'art sur la collection TRECVid SIN. La généricité de l'approche a été aussi validée avec des tests sur la collection KTH spécialisée pour la reconnaissance d'actions. Nous avons pu montrer que des vocabulaires de mots-trajectoire issus de la collection TRECVid générique, permettent d'obtenir de bons résultats sur cette autre base. A terme, des optimisations supplémentaires pourront encore améliorer les performances.

Remerciements

Nos descripteurs ont été obtenus en utilisant le centre de calcul MUST de l'Université de Savoie. Les programmes des classifieurs KNN et des mesures de performance ont été exécutés sur la plate-forme Grid'5000, en utilisant les outils logiciels du groupe IRIM (Indexation et Recherche d'Information Multimedia) du GdR ISIS, outils développés sous l'initiative INRIA ALADDIN avec support du CNRS, de RENATER et de plusieurs autres universités ou sources de financement (voir `https://www.grid5000.fr`).

Références

[1] J. Sivic and A. Zisserman, "Video google : a text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Oct., pp. 1470–1477 vol.2.

[2] S. Little, A. Llorente, and S. Rüger, "An overview of evaluation campaigns in multimedia retrieval," in *ImageCLEF*, Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo, Eds., vol. 32 of *The Information Retrieval Series*, pp. 507–525. Springer Berlin Heidelberg, 2010.

[3] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06 : Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[4] N. Ballas et al., "IRIM at TRECVID 2012 : Semantic Indexing and Instance Search," in *Proceedings of the workshop on TREC Video Retrieval Evaluation (TRECVID)*, Gaithersburg, MD, États-Unis, Nov. 2012, p. 12.

[5] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Comput. Vis. Image Underst.*, vol. 108, no. 3, pp. 207–229, Dec. 2007.

[6] M.-Y. Chen and A. Hauptmann, "Mosift : Recognizing human actions in surveillance videos," Tech. Rep. CMU-CS-09-161, Carnegie Mellon University, 2009.

[7] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories," in *Proceedings of the 2011 International Conference on Computer Vision*, Washington, DC, USA, 2011, ICCV '11, pp. 1419–1426, IEEE Computer Society.

[8] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *ICCV 2013 - IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 2013, IEEE.

[9] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[10] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar, "Trajectons : Action recognition through the motion analysis of tracked features," in *Workshop on Video-Oriented Object and Event Classification, ICCV 2009*, September 2009.

[11] N. Ballas, B. Delezoide, and F. Prêteux, "Trajectories based descriptor for dynamic events annotation," in *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, New York, NY, USA, 2011, J-MRE '11, pp. 13–18, ACM.

[12] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, June 2011, pp. 3169–3176.

[13] A. Gaidon, Z. Harchaoui, and C. Schmid, "Actom sequence models for efficient action detection," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2011, CVPR '11, pp. 3201–3208, IEEE Computer Society.

[14] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, Jun 1994, pp. 593–600.

[15] J. Y. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[17] D. Arthur and S. Vassilvitskii, "k-means++ : the advantages of careful seeding," in *SODA*, 2007, pp. 1027–1035.

[18] E. Yilmaz, E. Kanoulas, and J. A. Aslam, "A simple and efficient sampling method for estimating ap and ndcg," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2008, SIGIR '08, pp. 603–610, ACM.

[19] S.T. Strat, A. Benoit, and P. Lambert, "Retina enhanced sift descriptors for video indexing," in *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, 2013, pp. 201–206.

[20] S. T. Strat, *Analysis and Interpretation of Video Scenes through Collaborative Approaches*, Ph.D. thesis, University of Savoie, University Politehnica of Bucharest, Annecy, France, Dec. 2013.

[21] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions : A local svm approach," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, Washington, DC, USA, 2004, ICPR '04, pp. 32–36, IEEE Computer Society.