Sujet de thèse
2014-2017

Analyse, description et reconnaissance de contenu vidéo

Diane LINGRAND, lingrand@i3s.unice.fr
Frederic PRECIOSO, precioso@i3s.unice.fr

Les contenus vidéos sont présents dans un nombre toujours croissant de domaines, tant scientifiques que commerciaux. Citons par exemple les applications de TV interactive, diffusion de contenus numériques, vidéo à la demande (VOD), simulations, vidéo-conférences, etc. Les volumes de contenus vidéos actuels explosent, il suffit pour s'en convaincre d'analyser l'évolution du géant *Youtube* (en Mai 2010, plus de contenus vidéos étaient déposés sur le site de Youtube en 60 jours que tous les films réunis des trois plus grand studio américains sur 60 ans [Youtube2010], plus d'une heure de contenu vidéo est déposé actuellement chaque seconde sur le site de Youtube [Youtube2010, YoutubeNow]) d'où la nécessité de techniques spécifiques pour l'administration des bases de données vidéo, permettant la recherche et la navigation dans les contenus vidéo, tandis que les techniques actuelles ne permettent que la recherche par mots-clefs et la navigation en ligne. La recherche dans les domaines de la segmentation temporelle, du stockage, de la recherche et de la navigation dans les contenus vidéos s'est intensifiée sur les cinq dernières années. En effet, des avancées significatives ont été réalisées sur la dernière décennie sur la recherche d'information visuelle par le contenu. Il s'agit alors de détecter des caractéristiques visuelles dans les documents multimédia afin de les structurer ensuite en index à l'échelle du document. A cette étape de représentation des données, s'ajoute celle de la recherche proprement dite. Dans un article récent d'état de l'art sur les avancées en indexation et classification de contenus vidéos Hu et al. [HU11] placent comme premier verrou scientifique nécessitant des efforts de recherche : « la distinction efficace du mouvement de l'avant plan de celui de l'arrière plan, la détection et la reconnaissance des objets par leur mouvement et des évènements, la combinaison de caractéristiques visuelles statiques et de mouvement ainsi que la construction d'index intégrant l'information de mouvement ».
En matière d'analyse de l'apparence visuelle, l'extraction d'indices visuels pertinents dans les images, robustes à des variations géométriques et radiométriques a atteint une maturité certaine (voir [MS05, TM08] pour une description de ce type d'outils). Il en va de même côté classification où les nombreux développements en apprentissage statistique ont produit des outils puissants désormais utilisés dans des contextes très variés [V98, HTF01, STC02]. Citons pêle-mêle les séparateurs à vaste marge (SVM), les algorithmes de boosting, et les machines à noyaux comme les fers de lance de ces approches modernes par apprentissage automatique pour la classification.

Malgré tout, les tâches de reconnaissance visuelle se diversifient et se compliquent. Les catégories à classer et à retrouver dans les dernières campagnes internationales (TRECVID, Videolympics) pouvaient être des classes d'objets avec de grandes variabilités d'apparence, ou même représentant des concepts abstraits (*weather, disgust*), des scènes (*cityscape, airplane*), des évènements (*people protesting*) ou encore des actions (running, explosion). La représentation des données passe alors souvent par des structures complexes comme des séquences, des arbres ou des graphes pour tenter d'identifier l'information pertinente [HB07, GHS11, GHS12].

La nature des informations disponibles pour l'apprentissage évolue également. On est passé d'un cadre supervisé classique sur des données d'apprentissage vectorielles bien étiquetées à des contextes très différents comme, par exemple, le cas où les données ne sont que très partiellement étiquetées mais en nombre important. D'autre part, le volume des données toujours plus grand à traiter

change fondamentalement l'algorithmique des systèmes d'analyse et d'apprentissage. Le passage à l'échelle est un verrou important des applications actuelles.

Ce sujet de thèse se situe exactement à l'intersection des domaines de la Vision par ordinateur et de l'Apprentissage statistique et s'inscrit bien sûr dans des tâches de reconnaissance appliquées aux volumes de données réels à traiter aujourd'hui. Nous sommes convaincus que seule une coopération forte entre ces deux disciplines peut permettre l'émergence de solutions robustes aux problèmes de reconnaissance de formes posées par la massification actuelle des données vidéos.

1. **Analyse et description du contenu vidéo.** Si de nombreux descripteurs d'image peuvent être étendus aux données vidéo, il est nécessaire pour considérer l'information présente dans la dimension temporelle de ces données de proposer de nouveaux descripteurs, de nouvelles caractérisations, adaptés. Différentes pistes pourront être explorées tant au niveau de caractérisations globales de volume 2D+T vidéo à partir des résultats préliminaires intéressants de [GYT+09, MPV+12], qu'au niveau des caractérisations locales [ZPC11].
2. **Machine à noyaux pour l'apprentissage de catégories visuelles.** Afin de traiter des problèmes de catégorisation complexe, on propose ici d'explorer des représentations des données utilisant des fonctions noyaux adaptées. Outre la conception de fonctions noyaux dédiées au traitement des flux vidéo [ZPC11], nous souhaitons résolument intégrer des schémas d'apprentissage sur tout ou parties de ces noyaux structurés. Ainsi, des schémas basés sur le principe du Boosting ayant déjà montré des performances remarquables en classification d'images pourraient être étendus au contenu vidéo [LGP13].
3. **Indexation et recherche dans de très grandes bases.** Nous abordons systématiquement le problème du passage à l'échelle des méthodes développées. Nous proposons des structures d'indexation informatique de gros volumes pour la recherche rapide des plus proches voisins permettant de faire de la recherche de contenu dans des immenses bases, mais aussi des schémas rapides pour l'évaluation de fonctions noyaux sur des bases vidéos [GCP11, GPP11].

Informations

Direction : Diane Lingrand, F. Precioso.
Contact : lingrand@i3s.unice.fr, precioso@i3s.unice.fr
Lieu : Laboratoire I3S – UMR UNS-CNRS 7271 – Sophia Antipolis

Bibliographie

[GCP11] D. Gorisse, M. Cord, F. Precioso, SALSAS : Sub-linear Active Learning Strategy with Approximate k-NN Search, in Elsevier, Pattern Recognition, 44, 10-11 (October 2011), 2244-2254.

[GHS11] A. Gaidon, Z. Harchaoui, C. Schmid, A time series kernel for action recognition, BMVC 2011 – British Machine Vision Conference, Aug 2011.

[GHS12] A. Gaidon, Z. Harchaoui, C. Schmid, Recognizing activities with cluster-trees of tracklets, BMVC 2012 – British Machine Vision Conference, Sep 2012.

[GPP11] P-H Gosselin, F. Precioso, S. Philipp-Foliguet, Incremental Kernel Learning for Active Image Retrieval without Global Dictionaries, in Elsevier, Pattern Recognition, 44, 10-11 (October 2011), 2343-2357.

[GYT+09] X. Gao, Y. Yang, D. Tao and X. Li, Discriminative optical flow tensor for video semantic analysis, in Computer Vision and Image Understanding 113:372-383, 2009.

[HB07] Z. Harchaoui and F. Bach, Image classification with segmentation graph kernels, CVPR, 2007.

[HTF01] T. Hastie, R. Tibshirani, J. Friedman, The Element of Statistical Learning, Springer, Berlin, 2001.

[HU11] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, “A survey on visual content-based video indexing and retrieval,” Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 41, no. 6, pp. 797–819, 2011.

[LGP13] A. Lechervy, P.H. Gosselin, F. Precioso, Boosted kernel for image categorization, in Multimedia Tools and Applications (MTAP), (2013): 1-20 , January 19, 2013.

[MPV+12] V.F. Mota, E.A. Perez, M.B. Vieira, L.M. Maciel, F. Precioso, P.H. Gosselin, A tensor based on optical flow for global description of motion in videos, IEEE SIBGRAPI Conference on Graphics, Patterns and Images, 2012.

[MS05] K Mikolajczyk, C Schmid, A Performance Evaluation of Local Descriptors, in IEEE Transactions on Pattern Analysis and Machine Analysis, 2005.

[STC04] J. Shawe-Taylor and N. Cristianini, Kernel methods for Pattern Analysis, Cambridge University Press, 2004.

[TM08] T. Tuytelaars, K. Mikolajczyk, Local Invariant Feature Detectors: A Survey, Foundations and Trends in Computer Graphics and Vision, Vol. 3, nb 3, pp 177-280, 2008.

[V98] V.N. Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, 1998.

[Youtube2010] Youtube, Facts and Figures, "http://www.supermonitoring.com/blog/2010/05/17/youtube-facts-and-figures-history-statistics/," [Online].

[YoutubeNow] Youtube, OneHourPerSecond, "http://www.onehourpersecond.com/," [Online].

[ZPC11] S. Zhao, F. Precioso, M. Cord, Spatio-Temporal Tube data representation and Kernel design for SVM-based video object retrieval system, in Springer Journal on Multimedia Tools and APplications (MTAP), 55:(1), 105-125, October 2011.