IDIAP RESEARCH REPORT

INVESTIGATING TIME-SENSITIVE TOPIC MODEL APPROACHES FOR ACTION RECOGNITION

Romain Tavenard Remi Emonet
Jean-Marc Odobez

Idiap-RR-26-2013

JULY 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Investigating time-sensitive topic model approaches for action recognition

Romain Tavenard Rémi Emonet Jean-Marc Odobez

July 23, 2013

Abstract

In this paper, we present several attempts of using topic models for action recognition in videos. We show that time-sensitive topic models help recognizing actions when little training data is available. We also exhibit some limitations of these models when dealing with complex videos. New applications of these models in semi-supervised settings and the use of inherently discrimant models such as the MedLDA one are also considered.

1 Introduction

Action recognition is an important field of video processing. Its applications covers, among others, automatic annotation of videos, improved human-computer interaction and guidance in monitoring public spaces. Most state-of-the-art techniques for action recognition video documents rely on Bag-of-Word (BoW) representations. The latter are built from quantized spatio-temporal descriptors collected over long video segments [10, 14, 12, 13]. Such methods, however, do not encode the time information, although actions are characterized by strong temporal components. To address this issue and enhance action recognition performance, we investigate the use novel principled probabilistic methods (called topic models) for capturing the temporal relationships between characteristic sub-units of a given action. In a previous paper [17], we showed these models could help action recognition when little training information is available. In the following, we expose more experiments to better spot strong points and weaknesses of these models when used for action recognition.

2 Related work

Two main classes of features can be used for action classification [15]. First, global features can be computed on regions of interest (ROI) obtained from foreground subtraction or tracking techniques [24]. Second, local features can be extracted on a dense grid [20, 19] or computed around spatio-temporal interest points (STIP) [9].

Then, classification can be performed on these descriptors directly or after summarizing them into a new single feature, using a Bag-of-Word (BoW) approach [12] feeding a Support Vector Machine (SVM). These approaches lead to very competitive classification results, though they do not extract strong semantics from the data.

In order to include more temporal knowledge in the process, Hidden Markov Models (HMM) [23] or Conditional Random Fields (CRF) [16] can be used. Though, Markov models rely on the assumption that there is a single object in the scene performing a single action. Another way to add temporal information is to derive part-based models as in [13].

Finding recurrent meaningful activities in video data is especially relevant in the domain of video surveillance. Topic models, that originate from the text processing community, has become a relevant research direction to do so. Probabilistic Latent Semantic Analysis (PLSA) [6] or Latent Dirichlet Allocation (LDA) [2] were introduced to discover the dominant and semantically meaningful topics in large data collections through the co-occurrence analysis of words and allow to handle synonymy and polysemy of words. These models build on BoW representations. They have been used in various forms to discover human activities from sport [14], surveillance videos [21], accelerometers [7], or cell phone GPS [5]. Some attempts have been made to add supervision to these models in the field of text classification [8, 3, 25].

The inclusion of temporal information at different levels of the modeling has become an important research area [1, 22]. Recent evolutions of topics models [18, 11] integrate temporal information within the topics without enriching an exponential growth of the vocabulary as with n-grams.

This paper presents results that are complementary to those in [17]. Section 3 explores the use of parametric models that allow the time dimension of the data to be considered as continuous. In Section 4, basic models are evaluated in the case of simulatenous actions. Then, in Section 5, we compare several classification strategies ranging from simple voting scheme to more elaborate ones that make use of χ^2 kernel SVM to mix BoW information with motif one. We then present experiments that aim at using the unsupervised nature of our topic models so as to perform semi-supervised action recognition in Section 6. Finally, another set of experiments discusses the use of an already-published discriminative topic model (MedLDA [25]) for action recognition in videos (Section 7).

When not stated, experiments are performed on the KTH dataset, extracting STIP features that are later k-means quantized using 4,000 words. Parameters for STIP extraction are the ones used by Laptev in [10], except for Hollywood2 dataset for which the interest point detection threshold is set to 10^{-6} using the software of [10]. Computed features are histograms of flow (HoF) only for all datasets.

Method	KTH	Weizmann	Hollywood2	OlympicSports
BoW+SVM	**90.50%**	73.11%	**84.71%**	**62.00%**
PLSM	89.34%	80.64%	75.23%	51.44%
Parametric PLSM	89.34%	79.56%	74.39%	51.42%
HDLSM	89.46%	**82.79%**	73.12%	42.73%

Table 1: Classification accuracies for several models.

3 Model comparison

In this section, we present comparative performances of several models, namely HDLSM [4], PLSM [18] and a new parametric version of the latter.

In this parametric model, presence of a word in a motif is modelled using both a Gaussian on time and a uniform background noise that spreads all along the motif's time axis. In other words, $p(rt|w,z)$ is not modelled using a categorical distribution but rather using a continuous time model:

$$p(t_r|w,z) \propto C + \frac{1}{\sqrt{2\pi(\sigma_t^0+\sigma_t)^2}} \exp\left(-\frac{(t_r-\mu_t)^2}{2(\sigma_t^0+\sigma_t)^2}\right), \tag{1}$$

where σ_t^0 is an input parameter of the model that prevents from learning too sharp Gaussians that would hardly generalize, σ_t and μ_t are fit to the data and C corresponds to a uniform prior on relative times.

The framework used for classification is the same as the one used in [17]. In all cases, parameters are set so as to get 1 motif per class. Classification is performed using a simple voting scheme: a test video is assigned the label corresponding to the class on which its majoritarian motif was learned. Results are presented in terms of correct classification rate for KTH dataset using full training set (Table 1). When considering Hollywood2 [12] (training is performed using clean training data here) or OlympicSports [13] datasets, mAP metric is reported as suggested in original publications.

In all cases, PLSM and HDLSM model perform similarly and the parametric version of PLSM does not achieve better performance than its discrete time counterpart. Not surprisingly, when the amount of training data is large (which is especially true for OlympicSport and Hollywood2 datasets), SVM classification of BoW is very competitive. As noticed in [17], when little training data is available, as for Weizmann dataset, topic model approaches become interesting competitors. In the case of Hollywood2 and OlympicSports datasets, many spurious visual words are generated in the background, which makes the use of generative models that intend to explain all observed words unefficient. In these cases, BoW approach performs much better.

4 Simultaneous activities

Topic models are known to deal well with mixtures of topics occurring simultaneously. We hence built synthetic mixed activity datasets by mixing video

Method	KTH	Weizmann
BoW+SVM	**95.25%**	**99.73%**
HDLSM	82.29%	75.38%

Table 2: Mean average precision for simultaneous activities.

Number of topics per class	Voting scheme	SVM scheme	Mixed SVM scheme
1	**89.34%**	87.14%	87.95%
2	**88.64%**	86.44%	88.18%
5	**89.69%**	85.98%	88.41%
10	**88.64%**	83.55%	86.33%

Table 3: Classification accuracies for several classification schemes.

samples from 2 different activity classes. These synthetic datasets are buit from the KTH and Weizmann one.

In order to take into account rankings from several video classes, mean average precision (mAP) is used to assess the quality of the class ranking returned by standard BoW approach compared to HDLSM one. Results presented in Table 2 show that BoW significantly outperforms HDLSM for this task, showing performance that is close to perfection on these 2 datasets.

5 Classification scheme and number of motifs

In this section, the same framework is used to fit motifs to the data and standard PLSM model is considered. Three classification schemes that operate at the output of this fitting step and hence make use of the motif probabilities in the documents are compared. Firstly, the "voting scheme" is the same as the one used in the previous section, that is the clasification decision is made following the formula:

$$C(d) = \underset{C_i \in \{C_1,\ldots,C_N\}}{\arg\max} \sum_z p(y = C_i|z)p(z|d), \tag{2}$$

where $p(y = C_i|z)$ terms are either 1 (if motif z has been learned on class C_i) or 0 (otherwise). Secondly, the "SVM scheme" consists in learning a χ^2 kernel SVM classifier on the $p(z|d)$ distribution and using this classifier to make the decision for a new video on which motifs are inferred. Finally, the "mixed SVM scheme" consists in merging informtaion from both the $p(z|d)$ distribution and the BoW using a multi-channel χ^2 kernel. In all cases, results are reported with respect to the number of motifs learned per class, as one could expect that when using SVM, more motifs could help make the decision. Results are presented for KTH dataset using full training set with varying number of topics (Table 3) and 2 other datasets with 1 topic learned per class (Table 4).

In all cases, the voting scheme appears to perform the best, showing that the use of a classifier at the output of our system is useless, as learned motifs

Classification scheme	Weizmann	Hollywood2
Voting scheme	**80.64%**	**75.23%**
SVM scheme	53.76%	58.55%
Mixed SVM scheme	55.91%	61.33%

Table 4: Classification accuracies for several classification schemes on Weizmann and Hollywood2 datasets.

are sufficiently tied to a given class for voting to be efficient.

6 Supervision

In this section, we compare supervised and semi-supervised approaches for PLSM using the full KTH dataset. In the semi-spervised approach, a single model made of 30 motifs (this number has been chosen for improved performance) is learned using the full training set and supervision is introduced by learning a SVM on the motif probability tables. For each labelled set size, classification accuracy results are reported for the semi-supervised approach together with those for a fully supervised setup using only the considered labelled set. Results are presented in Fig. 1.

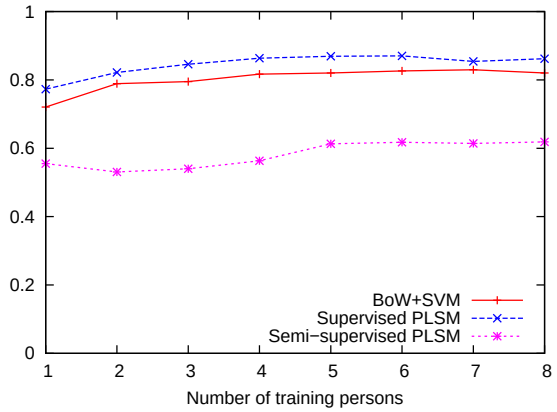
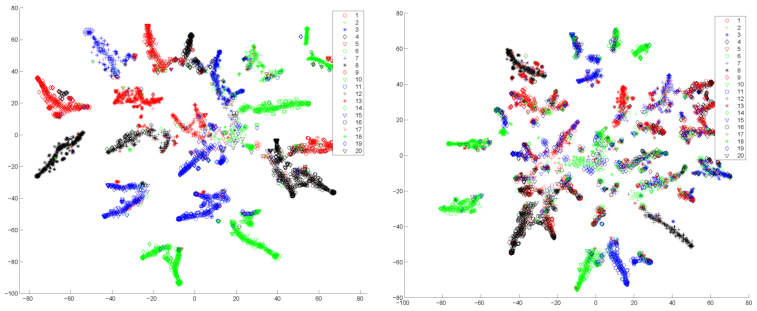


Figure 1: Classification accuracies for several supervision strategies on subparts of the KTH dataset.

Poor performance of the semi-supervised approach indicates that learning motifs without any prior on their class association is weaker than learning motifs on smaller training sets (as for the baseline, the unlabelled data is not used for training) in a supervised manner.

7 MedLDA

MedLDA [25] is a discriminative topic model built on top of LDA in which topics are learnt so as to be both representative of the documents in the training collection and efficient at discriminating between classes. Fig. 2 illustrates this on a text corpus, showing that documents of the same class tend to be close in the topic space.



(a) (b)

Figure 2: 2d embedding of the discovered latent representations on the 20newsgroup dataset by the MedLDA (a) and standard unsupervised LDA (b). The 2d embeddings are achieved with the t-SNE (t-Distributed Stochastic Neighbor Embedding) method. This figure comes from [25].

The principle of MedLDA is to define a *partially* generative model on (θ, z, W) as in LDA and to apply max-margin principle for the classification (link between Z_d and Y_d) part (see Fig. 3).

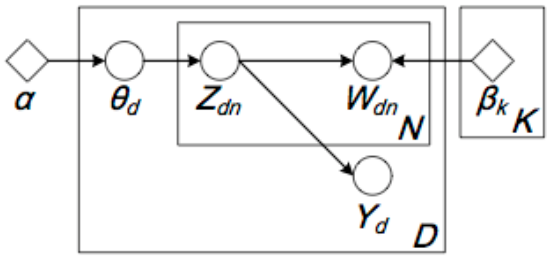


Figure 3: Plate notation of the MedLDA model.

The updates of topic probabilities in a document d is done according to:

$$\gamma_d = \alpha + \sum_{n=1}^{N} \phi_{dn}, \tag{3}$$

Setup	20 Newsgroup (Text corpus)	KTH (Video corpus)
During training	99.8%	95%
On training data	96.8%	80%
On test data	79.6%	75%

Table 5: MedLDA classification performance.

where

$$\phi_{dn} \propto \exp\left(\mathbb{E}[\log\theta_d|\gamma_d] + \log p(w_{dn}|\beta) \quad \underbrace{+\frac{1}{N}\sum_{y\in\mathcal{C}}\hat{\mu}_d^y\mathbb{E}[\eta_{y_d} - \eta_y]}_{\text{Max-margin optimization related term}}\right). \quad (4)$$

The important point here is that $\hat{\mu}_d$ will be non-zero for documents that lie on the decision boundary (*i.e.* Support Vectors) so as to push them to the correct side of the boundary. Though, when inferring topic probabilities for a new document to be classified, class label is unknown and this term then disappears, which is likely to alter topic probabilities in a non-neglectable way. Hence, the key for this method to work properly is that the model should converge sufficiently enough so that these terms become almost useless in the optimization process.

Table 5 presents MedLDA classification performance (for MedLDA with 30 topics) obtained by three means. First (denoted as *during training*), we report classification performance that is obtained during the training process (in this case, the topic probabilities are computed using the SVM term). Then (denoted as *on training data*), we provide classification performance when inferring topic probabilities on the same data (but in this case, the SVM term is not present any more in the formulas). Finally (denoted as *on test data*), we report classification performance on test data.

While for the text corpus, performance stays very high between rows 1 and 2 of our table, in the case of video data, the phenomenon that was discussed earlier proved to happen: for a given set of documents, when labels are removed, topic distributions in documents are altered and the performance drops. It is then not surprising that when applied on test data, the performance is weaker than that of our PLSM approach presented above.

8 Conclusion

In this paper, we presented experiments that complements those presented in [17] regarding the use of topic models for action recognition in videos. New experiments include application to more recent datasets such as Hollywood2 or Olympic Sports ones or synthetic mixed activity datasets. We also study in

more details some implementation details such as the use of a classifier at the output of our system instead of a simple voting scheme or the number of topics to be used. Then, a semi-supervised alternative is presented and compared to our fully supervised setting. Finally, we experimentally demonstrate that for such complex data in which reproducibility of word generation in video documents is not straightforward, state-of-the-art discriminative topic models such as the MedLDA one do not perform as well as our proposed approach.

Acknowledgments

This work was partially funded by SNSF (Swiss National Science Foundation) through the PROMOVAR project.

References

[1] David M. Blei and John D. Lafferty. Dynamic topic models. In *International conference on Machine learning*, pages 113–120, 2006.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] D.M. Blei and J.D. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128, 2008.

[4] R. Emonet, J. Varadarajan, and J.-M. Odobez. Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2011.

[5] Katayoun Farrahi and Daniel G. Perez. What did you do today?: discovering daily routines from large-scale mobile data. In *ACM international conference on Multimedia*, pages 849–852, 2008.

[6] Thomas Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[7] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *International conference on Ubiquitous computing*, pages 10–19, 2008.

[8] S. Lacoste-Julien, F. Sha, and M.I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems*, 21, 2008.

[9] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, pages 432–439, 2003.

[10] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2008.

[11] Jian Li, S. Gong, and T. Xiang. Discovering multi-camera behaviour correlations for on-the-fly global activity prediction and anomaly detection. In *IEEE International Workshop on Visual Surveillance*, 2009.

[12] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[13] J.C. Niebles, C.W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision*, pages 392–405, 2010.

[14] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79:299–318, 2008.

[15] Ronald Poppe. A survey on vision-based human action recognition. *Image Vision Computing*, 28(6):976–990, 2010.

[16] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *International Conference on Computer Vision*, volume 2, pages 1808–1815, 2005.

[17] Romain Tavenard, Rémi Emonet, and Jean-Marc Odobez. Time-sensitive topic models for action recognition in videos. In *International Conference on Image Processing*, 2013.

[18] Jagannadan Varadarajan, Rémi Emonet, and Jean-Marc Odobez. Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. In *British Machine Vision Conference*, 2010.

[19] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, 2011.

[20] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

[21] Xiaogang Wang, Xiaoxu Ma, and W.E.L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:539–555, 2009.

[22] Xuerui Wang, Andrew Mccallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *IEEE International Conference on Data Mining*, 2007.

[23] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 379–385, 1992.

[24] Bangpeng Yao and Li Fei-Fei. Action recognition with exemplar based 2.5d graph matching. In *European Conference on Computer Vision*, 2012.

[25] Jun Zhu, Amr Ahmed, and Eric P. Xing. MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research*, 13:2237–2278, 2012.