Time-Sensitive Topic Models for Action Recognition in Videos

Romain Tavenard, Rémi Emonet, Jean-Marc Odobez

TIME-SENSITIVE TOPIC MODELS FOR ACTION RECOGNITION IN VIDEOS

Romain Tavenard, Remi Emonet, Jean-Marc Odobez

Idiap Research Institute, Martigny, Switzerland
{romain.tavenard,remi.emonet,odobez}@idiap.ch

ABSTRACT

In this paper, we postulate that temporal information is important for action recognition in videos. Keeping temporal information, videos are represented as word×time documents. We propose to use time-sensitive probabilistic topic models and we extend them for the context of supervised learning. Our time-sensitive approach is compared to both PLSA and Bag-of-Words. Our approach is shown to both capture semantics from data and yield classification performance comparable to other methods, outperforming them when the amount of training data is low.

1. INTRODUCTION

Action recognition is key for many tasks such as automatic annotation of videos, improved human-computer interaction and guidance in monitoring public spaces. As the amount of available videos from different sources (from raw personal videos to more professional content) has dramatically increased in the last few years, new propositions are needed to organize this new data.

Many recent state-of-the-art techniques for action recognition in naturalistic and unconstrained video documents such as movies or broadcast data rely on Bag-of-Word (BoW) representations built from quantized spatio-temporal descriptors extracted at Spatio-Temporal interest points (STIP) or on a dense grid and collected over long video segments [1, 2, 3, 4]. Such methods, however, often suffer from two severe and related drawbacks:

- the time information is often discarded, although actions are characterized by strong temporal components;
- activities within the same video segments are mixed in such BoW representation, which can negatively affect recognition algorithms that are based on these.

To address these issues 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. The principle is illustrated in Fig. 1. A video clip (on top) is represented as a STIP word×time document. The idea is that when a person performs an action, it leaves a temporal trace in the document, where the trace is formed by a set of characteristic words occurring at a given time after the start of the action. In the Figure, such a trace for a hand-clapping sequence is highlighted with color words Importantly, note that other words (e.g., the ones generated by other activities in the scene) can simultaneously occur.

The aim is thus, from training video clips/documents, to automatically recover the relevant temporal patterns (or motifs, represented as probability tables, see right of Fig. 1) associated with an action, and further automatically identify their occurrences in test sequences. To this end, we leverage on recent work that automatically discover temporal motifs from word×time documents [5] in an unsupervised fashion. Applied to large volume of video-surveillance data, such a method has shown to capture not only the co-occurrence between words but also the order in which they occur.

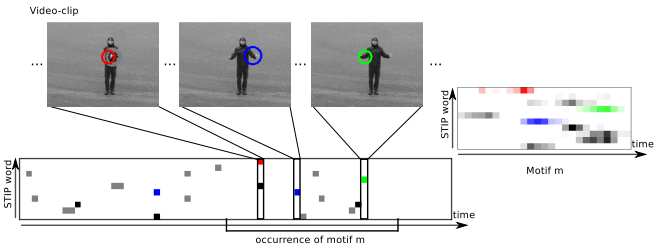


Fig. 1. Schematic view of our approach. Videos are seen as word×time tables and extracted motifs aim at capturing the temporal order of visual words inside actions (see text for more details).

In this context, our main contributions is to investigate the video action recognition task from STIP word, by automatically identifying meaningful and interleaved temporal patterns with temporal support longer than those of STIP, and addressing the above-mentioned challenges. Furthermore, we extend the model of [5] to learn motifs in a supervised framework. As such, through the use of supervised time-sensitive topic models, our work proposes an alternative to recent pieces of work that make use of temporal information to achieve this task [4]. Note as well that contrary to state-based temporal models such as HMMs or CRFs, our approach can deal with interleaved activities as well as with activities that are made of several simultaneous action sub-units.

We show that our method achieves very competitive classification performance, especially when only little training data is available. In addition, we show that it also extracts nice semantic patterns from the data and models well action primitives.

This paper is organized as follows. Section 2 reviews related works. Section 3 presents the basics of the topic model proposed in [5]. It is extended for supervised action recognition from video, as exposed in Section 4. Experiments are shown in Section 5.

2. PREVIOUS WORK

In this paper, we are interested in temporal activity modeling and action recognition. We are particularly interested in the case of action recognition and mining in videos but the relevant work cover more than these two.

2.1. Action Recognition in Videos

Moeslund *et al.* [6] suggest to divide the task of vision-based action classification into three levels of abstraction. The first level is

referred to as *action primitive* and corresponds to an atomic movement. Based on action primitives, *actions* can be derived that are coherent sequences of action primitives. Finally, *activities* are defined as larger scale events that depend on the context. Authors illustrate these concepts in the case of a tennis match where "forehand", "backhand" and "run left" are examples of action primitives. An action is then the set of action primitives needed to return a ball and the activity is "playing tennis".

According to [7], two classes of features can be used for action classification. On the one hand, global features can be computed on regions of interest (ROI) obtained from foreground subtraction or tracking techniques [8]. On the other hand, local features can be extracted on a dense grid [9, 10] or computed around spatio-temporal interest points (STIP) [11]. Classification can be performed on these descriptors directly or after summarizing them into a new single feature, using a Bag-of-Word (BoW) approach [3]. Usually, large margin methods like Support Vector Machines (SVM) are used for classification. These approaches lead to very competitive classification results, though they do not extract strong semantics from the data.

However, these techniques use no or very little temporal knowledge, which can be improved by using temporal state models such as Hidden Markov Models (HMM) [12] or Conditional Random Fields (CRF) [13]. Markov models have been used to capture temporal information but unfortunately they usually use global features and rely on the assumption that there is a single object in the scene performing a single action. Niebles *et al.* [4] have derived a temporal formulation of part-based models introduced in [14] for action recognition. An open challenge is to both model temporal information and handle mixture of actions.

2.2. Topic Models for Activity Mining in Videos

The task of activity mining (finding recurrent meaningful activities) in video data has attracted a lot of interest, particularly in the domain of video surveillance. In many cases, it relies on the extraction of streams of features from a camera, and involves the modeling of the temporal evolution and interactions of these features streams to infer some activity category.

Recently, the design of probabilistic Bayesian models called *topic models* has become a relevant research direction to discover recurrent patterns in sensor data. These models originate from the text processing community. Topic models such as Probabilistic Latent Semantic Analysis (PLSA) [15] or Latent Dirichlet Allocation (LDA) [16] build on top of BoW. They 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. They have been used in various forms to discover human activities from sport [2], surveillance videos [17], accelerometers [18], or cell phone GPS [19]. Some attempts have been made to add supervision to these models in the field of text classification [20, 21].

While initially designed to handle static data, the inclusion of temporal information at different levels of the modeling has become an important research area [22, 23]. For instance [23] proposed to model topics on n-gram words, while dynamic topic models [22] capture the evolution of topics in a sequentially organized corpus of documents. Attempts to model temporal information have been made: [24] introduces a temporal model on scene-level behaviors, and [25] models each topic as distribution on feature$\times$time words, but under the strong assumption of temporally aligned clips. Still, overall the temporal modeling and segmentation of activities using topic models was seldom addressed until recently.

Recent evolutions of topics models [26, 27] integrate temporal information within the topics without enriching an exponential growth of the vocabulary as with n-grams. [26] models topics as temporal patterns called "motifs" and can be seen as a probabilistic version of the methods presented in previous section. In [26], the method models and automatically finds both what the recurrent motifs (topics) are and when they occur in each of the documents from the learning set. These approaches have been successfully applied to find recurrent activities in traffic videos and surveillance videos, for which simple features are used, such as quantized optical flow.

Topic models such as PLSA or LDA naturally handle mixtures but are seldom used for the action classification. This is probably due to the apparent mismatch between the unsupervised nature of topic models and the supervised nature of the classification task. They are used in [2] with success and allow to properly handle mixed activities. Note however that in [2], the approach is fully unsupervised and hence performs action clustering rather than action recognition. Finally, in this approach, temporal information is not modeled. This restriction is acceptable as long as the actions are relatively short but becomes more problematic for longer actions.

To summarize this related work review, BoW approaches have shown very competitive results in the field of action recognition, though allowing little semantic analysis. Topic models, that would allow this kind of analysis as it has been shown in other domains, have not been much investigated for this application up to now. Moreover, to the best of our knowledge, supervised time sensitive topic models have not been investigated at all for action recognition up to now.

3. HIERARCHICAL DIRICHLET LATENT SEQUENTIAL MOTIFS (HDLSM)

In this paper, we build upon the HDLSM topic model that was first introduced in [5]. This generative model relies on the extraction of motifs that encapsulate the temporal information of the data. It is able to automatically find both the underlying number of motifs needed to model a given set of documents and the number of motif occurrences in each document (which includes their temporal locations), as shown in Figure 2a. In this section, we will present a brief overview of the model and its inference process, as this is key for our application.

3.1. Model

The HDLSM model takes as input a set of temporal documents. Each temporal document is represented as a table of counts that informs, for each pair (w, t), about the amount of presence of word w at time instant t. HDLSM generative process is globally as follows:

- Generate a list of motifs, each motif Φ_k being a 2D probability map indicating how likely it is that word w occurs at relative time rt after the beginning of the motif.
- For each document j, generate a list of occurrences, each occurrence having a starting time ost and a associated motif k.
- For each observation i in document j:
  - Draw an occurrence from the list,
  - Draw (w, rt) from the associated motif,
  - Generate the observation as $(w, ost + rt)$.

The advantage of such probabilistic generative models is that we can reverse the process and derive algorithm to learn all the model parameters from observed data.

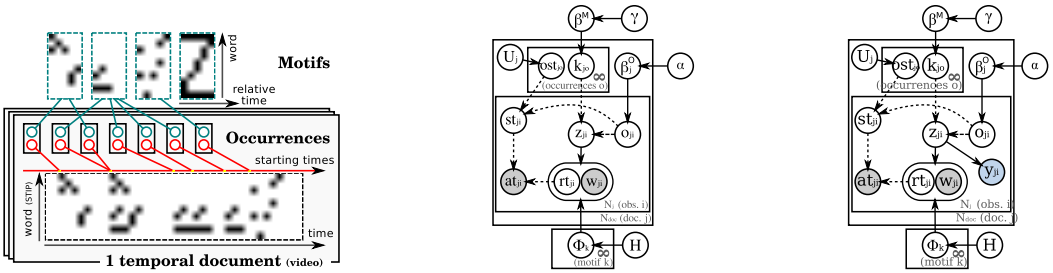


Fig. 2. HDLSM model. (a) presents a schematic view of how the model is used while (b) details the model with developed Dirichlet processes (using stick-breaking convention) and (c) presents its supervised counterpart.

3.2. Joint Inference of Motifs and Occurrences

A temporal document can be seen as a list of pairs (w_{ji}, at_{ji}) where the word w_{ji} appears at time at_{ji} in document j, forming its i^{th} observation. Moreover, motifs are represented as probabilistic maps, denoted as Φ. Each map is drawn from H, that is defined as a Dirichlet. This models makes intensive use of Dirichlet Processes (DP) to model the possibly infinite number of motifs and occurrences.

The global distributions over motifs β_M (with concentration parameter γ) and U_j are used to associate motif indices k_{jo} and starting times ost_{jo} to each occurrence, while the document specific distribution β_{jo} (with concentration parameter α) is used to sample the occurrence o_{ji} associated to each observation. Then, knowing o_{ji} and all occurrences, it is straight-forward to deduce the motif z_{ji} and starting time st_{ji} associated to an observation.

To learn the parameters of the model, a Gibbs sampling is applied, in which it is sufficient to re-sample o_{ji} for each observation and k_{jo} and ost_{jo} for each occurrence. Other variables are either integrated out or deduced, when a deterministic relation holds.

3.3. Inference on test documents

Let us assume that one has already learned a set of motifs that he knows to be well-adapted to the data he is considering. Then, given a new temporal document, fitting this set of motifs to the document can be done by sampling the occurrences alone. At each iteration, the algorithm will update the set of occurrences, their starting times and associated motifs, keeping the motifs (probability maps, Φ) unchanged. At the output of this process, one gets a set of occurrences that enable reconstruction of the temporal document using the fixed motif set.

4. SUPERVISED HDLSM FOR ACTION RECOGNITION IN VIDEOS

In this section, we present our supervised HDLSM model that we use for action recognition. We will detail the whole process, from video description to actual action recognition, including learning the model presented in Figure 2c.

4.1. Word×time document generation

In order to apply HDLSM model to our problem, we first need to turn video sequences into word×time documents that store, for each word at each time instant, its amount of presence in the video. To do so, we start by computing local features at salient points in the 2D+t domain, as using interest point detectors has shown to achieve better performance on the datasets we use than dense sampling [9]. This is done following the method of [1], which computes histograms of optical flow (HoF) in cuboids around interest points extracted using a space-time extension of the Harris operator. Once these features are computed, they can be quantized using a codebook of visual words, which is learned on a subset of feature points using k-means algorithm. The quantization step then consists in assigning to a given feature point the quantization index corresponding to its closest centroid (in terms of Euclidean distance). In our experiments, we used $k = 4,000$ as in [1]. Finally, a temporal document similar to the one presented in Figure 2a is built that stores at position (i, t) the number of features that were quantized into quantization index i at time t.

4.2. Learning temporal motifs using HDLSM

Following the work from [28], supervision is added to our model by adding a step to our generative process that consists in generating a class label y for each word with probability $p(y|z)$. In our case, the distribution of $p(y|z)$ values is set to a delta function, which is equivalent to learning motifs in a per-class fashion, using the method presented in Section 3.2. This results in as many sets of motifs as action classes. Note that the number of motifs extracted may vary from one class to another. However, association of motifs to classes (*i.e.* $p(y|z)$ distributions) are stored as they will be key for the recognition step described in the following.

4.3. Action recognition using learned motifs

When a new video comes in, its corresponding temporal document is generated in the exact same way as explained above. Then, inference is performed to find out, given the fixed set of motifs, what occurrences of motifs explain best the given temporal document d. This inference step gives us a probability map $p(z|d)$ on which the decision can be performed as a voting scheme:

$$C(d) = \underset{C_i \in \{C_1,\dots,C_N\}}{\arg\max} \sum_z p(y = C_i|z)p(z|d). \qquad (1)$$

5. EXPERIMENTAL VALIDATION

5.1. Experimental details

Datasets: We evaluate our approach on KTH [29] and Weizmann [30] datasets. KTH dataset is made of short video clips from 25 persons performing 6 actions and captured from 4 view points. Weizmann one contains 10 actions performed by 9 persons.
Evaluation protocol: We follow the standard experimental protocols. For KTH, the 25 persons are split in a training, validation, and

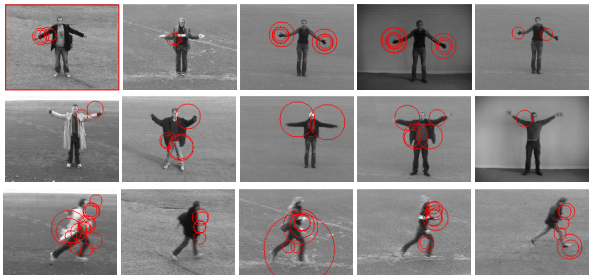
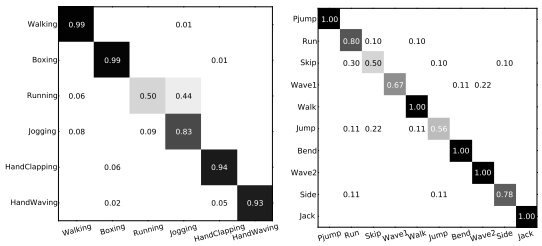


Fig. 3. The five strongest occurrences for motifs associated to "HandClapping" (row 1), "HandWaving" (row 2) and "Running" (row 3). For each occurrence, we show the image in the middle of the temporal segment associated to the occurrence, as well as the STIP features contributing to the occurrence in that frame. Note that almost all these occurrences happen in video clips from the right action category associated with the motif, the only exception being for class "HandClapping" for which one occurrence from class "HandWaving" (marked with red border) is retrieved.



(a) KTH (b) Weizmann

Fig. 4. HDLSM confusion matrices for both datasets.

Dataset	BoW+SVM	PLSA	HDLSM
KTH	90.50%	88.30%	89.46%
Weizmann	73.11%	77.41%	82.79%

Table 1. Comparative results in terms of classification accuracy. KTH dataset is used with large training set.

test sets [31]. Given our algorithms where no parameter tuning is necessary, we used both the training and validation sets for learning the motifs and the SVM in the baseline (as done in [1]), except for Figure 5 that shows the results evolution as a function of the number of people used for training. All classification accuracy results are computed on the test set. In the Weizmann dataset we use a standard Leave-One-Person-Out cross-validation method. For all methods, the same parameters are used for both datasets. In particular, the temporal length of an HDLSM motif is set to be 25 frames.

Baselines: We compare ourselves to a traditional BoW approach [1] that uses a χ^2 kernel SVM classifier for classification decision. From the class of methods that do not track people and that rely on STIP features, it is still one of the best performing methods. We also present results using the PLSA approach, since it is a model that also identify topics through co-occurrence analysis, but discards temporal information. This should highlight the actual benefit of using a time-sensitive model like HDLSM. To provide a fair comparison, the PLSA document are built using the same temporal support as HDLSM motif. Moreover, since PLSA does not model selection, the number of topics was set to the same value as the number of motifs in the HDLSM approach. In the experiments presented here, one motif is learned for each class.

5.2. Results

Qualitative results. One possibility to understand the semantics captured by our motifs is to show their strongest occurrences (as measured by the number of associated words) on the test set. This is illustrated in Figure 3 on the KTH dataset for motifs from 3 different actions by displaying one image of each occurrence. As can be seen, for a given motif, images are time-aligned (i.e. they are located at the same moment in the corresponding action), demonstrating our model's ability to grasp the typical temporal structure of actions.

Quantitative results. Results are presented in Table 1. They show that our approach is slightly better than PLSA, showing the interest of the temporal information when forming topics. Moreover, the performances of our method are very close to those of the BoW approach on the KTH dataset, while being much better on the Weizmann one. This is due to the ability of our method to deal with small training sets. This is further demonstrated in Figure 5 that shows that even on the KTH dataset, when small training sets are considered, our method outperforms the baseline.

Confusion matrices obtained for our method on both datasets are presented in Figure 4. On the KTH dataset, most of the confusion comes from the difficult distinction between "Running" and "Jogging" classes. On the Weizmann one, the class that leads to poorest results is the "Skip" one, which is due to the complexity of this class. Note however that the decomposition of "Skip" action into "Run" and "Jump", that is observed here, makes sense.

6. CONCLUSION

In this paper, we presented the use of supervised time sensitive topic models for action recognition in videos. We use existing time-sensitive unsupervised models and extend them by adding supervised classification capabilities. We showed through experiments that our method is able to both extract semantic motifs from the data and outperform standard approaches of the field when little training data is available.

Acknowledgements

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

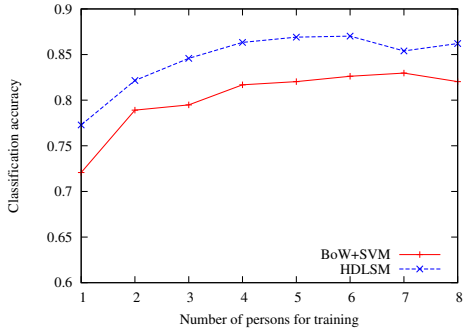


Fig. 5. Classification performance and training set size. Results reported here are computed using 1 to 8 persons for training on KTH.

7. REFERENCES

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

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

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

[4] 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*, 2010, pp. 392–405.

[5] 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.

[6] T.B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[7] Ronald Poppe, "A survey on vision-based human action recognition," *Image Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

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

[9] 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.

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

[11] Ivan Laptev and Tony Lindeberg, "Space-time interest points," in *International Conference on Computer Vision*, 2003, pp. 432–439.

[12] 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*, 1992, pp. 379–385.

[13] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in *International Conference on Computer Vision*, 2005, vol. 2, pp. 1808–1815.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[15] Thomas Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

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

[17] 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*, vol. 31, pp. 539–555, 2009.

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

[19] 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*, 2008, pp. 849–852.

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

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

[22] David M. Blei and John D. Lafferty, "Dynamic topic models," in *International conference on Machine learning*, 2006, pp. 113–120.

[23] 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.

[24] T. Hospedales, S. Gong, and Tao Xiang, "A markov clustering topic model for mining behavior in video," in *International Conference on Computer Vision*, 2009.

[25] T. A Faruquie, P. K Kalra, and S. Banerjee, "Time based activity inference using latent dirichlet allocation," in *British Machine Vision Conference*, 2009.

[26] 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.

[27] 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.

[28] A. Krithara, M.R. Amini, J.M. Renders, and C. Goutte, "Semi-supervised document classification with a mislabeling error model," in *Proceedingsof the European Conference on Information Retrieval*, 2008.

[29] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *IEEE International Conference on Pattern Recognition*, 2004, vol. 3, p. 32–36.

[30] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[31] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: A local svm approach," in *IEEE International Conference on Pattern Recognition*, 2004.