Probabilistic Latent Sequential Motifs: Discovering temporal activity patterns in video scenes

Jagannadan Varadarajan[12]
varadarajan.jagannadan@idiap.ch
Rémi Emonet[1]
remi.emonet@idiap.ch
Jean-Marc Odobez[12]
odobez@idiap.ch

[1] Idiap Research Institute
CH-1920 Martigny, Switzerland

[2] École Polytechnique Fédéral de Lausanne
CH-1015, Lausanne, Switzerland

Abstract

This paper introduces a novel probabilistic activity modeling approach that mines recurrent sequential patterns from documents given as word-time occurrences. In this model, documents are represented as a mixture of sequential activity motifs (or topics) and their starting occurrences. The novelties are threefold. First, unlike previous approaches where topics only modeled the co-occurrence of words at a given time instant, our topics model the co-occurrence and temporal order in which the words occur within a temporal window. Second, our model accounts for the important case where activities occur concurrently in the document. And third, our method explicitly models with latent variables the starting time of the activities within the documents, enabling to implicitly align the occurrences of the same pattern during the joint inference of the temporal topics and their starting times. The model and its robustness to the presence of noise have been validated on synthetic data. Its effectiveness is also illustrated in video activity analysis from low-level motion features, where the discovered topics capture frequent patterns that implicitly represent typical trajectories of scene objects.

1 Introduction

Immense progress in sensor and communication technologies has led to the development of devices and systems recording multiple facets of daily human activities. This has resulted in an increasing interest for research on the design of algorithms capable of inferring meaningful human behavioral patterns from the data logs captured by sensors, simultaneously leading to new application opportunities. The surveillance domain is a typical example. In scenes such as those illustrated in Fig.1, one would like to automatically discover the typical activity patterns, when they start or end, or predict an object's behavior. Such information can be useful in its own right, e.g. to better understand the scene content and its dynamics, or for higher semantic level analysis, to manually define the real camera activities, provide context for other tasks (e.g. object tracking) or spot abnormal situations which could for instance be used to automatically select the camera streams to be displayed in control rooms of public spaces monitored by hundreds of cameras.

BMVC 2010 doi:10.5244/C.24.117

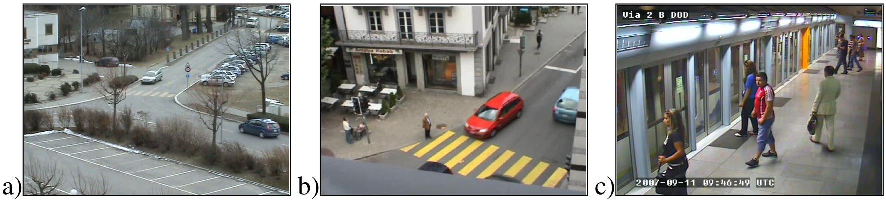


Figure 1: Surveillance scenes.a) Far Field b) Traffic Junction c) Metro

In this paper, we propose an unsupervised approach based on a novel graphical topic model to discover dominant activity patterns from sensor data logs represented by word×time count documents. The novelties are that i) the estimated patterns are not merely defined as static distribution over words but also incorporate the temporal order in which words occur; ii) the approach handles data resulting from the temporal overlap between several activities; iii) the model allows to automatically estimate the starting times of the activity patterns. Although the approach is general, we demonstrate its application to the discovery of dominant activities in surveillance scenes.

Most approaches addressing activity discovery are object-centered in which, objects are first detected, then tracked and their trajectories are used for further analysis [5, 10, 11, 13]. Tracking-based approaches provide direct object-level semantic interpretation, but are sensitive to occlusion and tracking errors especially in crowded scenes where multiple activities occur simultaneously. To avoid these problems, approaches relying on low-level features like location and velocity and their statistics rather than tracks have been proposed [8, 12, 15, 16]. Among them, topic models like pLSA [6] or LDA [2] have been shown to be promising approaches to discover scene level activity patterns through co-occurrence analysis of low-level features and detect abnormal events [8, 12, 14]. For instance, [14] uses a hierarchical variant of LDA to extract atomic actions and interactions in traffic scenes, while [8] relies on hierarchical pLSA to identify abnormal activities and repetitive cycles.

One important challenge is the actual modeling of temporal information: by relying only on the analysis of unordered word co-occurrence within a time window, most topic models fail to represent the sequential nature of activities. For example, in traffic scenes, people wait at zebras until all vehicles have moved away before crossing the road, giving rise to a temporally localized and ordered set of visual features. Using a "static" distribution over features to represent this activity will not allow to distinguish it from an abnormal situation where a person crosses the road while vehicles are still moving.

Few approaches have been made to incorporate temporal information in topic models. However, this was done either to represent single word sequences [4], or at the high level, i.e. by modeling the dynamics of topic distributions over time [1, 7]. For instance, [7] introduces a Markov chain on scene level behaviors, but each behavior is still considered as a mixture of unordered (activity) words. Both cases cannot address the modeling of multiple activities defined by temporal word patterns and that can happen concurrently and independently (e.g. motion of pedestrian and cars), a common situation in surveillance data. An attempt was made in [9], which modeled topics as feature×time temporal patterns, trained from video clip documents where the timestamps of the feature occurrences relative to the start of the clip were added to the feature. However, in this approach, the same activity has different word representations depending on its temporal occurrence within the clip, which prevents the learning of consistent topics from the regularly sampled video clip documents. To solve this issue of activity alignment w.r.t. the clip start, [3] manually segmented the videos so that the start and end of each clip coincided with the traffic signal cycles present in the scene.

This method has two drawbacks: first, only topics synchronized with respect to the cycle start can be discovered. Second, such a manual segmentation is time consuming and tedious. Our model allows to address all the above issues.

The rest of the paper is organized as follows. Section 2 describes our generative model. Section 3 demonstrates its validity and properties on synthetic data, while Section 4 presents its application to videos from two different scenes. Section 5 concludes the paper.

2 Probabilistic Latent Sequential Motif Model

In this paper, we propose a probabilistic model that explicitly describes the starting times of topics within a document as well as the temporal order in which words occur within a topic. In this section, we first introduce the notations and provide an overview of the model, and then describe with more details the generative process and the EM steps derived to infer the parameters of the model.

2.1 Model overview

Figure 2a illustrates how documents are generated in our approach. Let D be the number of documents d in the corpus, each having N_d words and spanning T_d discrete time steps . Let $V = \{w_i\}_{i=1}^{N_w}$ be the vocabulary of words that can occur at any given instant $t_a = 1,..T_d$. A document is then described by its count matrix $\mathrm{n}(w,t_a,d)$ indicating the number of times a word w occurs at the absolute time t_a within the document. These documents are generated from a set of N_z topics $\{z_i\}_{i=1}^{N_z}$ assumed to be temporal patterns $p(w,t_r|z)$ with a fixed maximal duration of T_z time steps (i.e. $0 \leq t_r < T_z$), where t_r denotes the relative time at which a word occurs within a topic, and that can start at any time instant t_s within the document[1]. In other words, qualitatively, documents are generated in a probabilistic way by taking the topic patterns and reproducing them at their starting positions within the document, as illustrated in Fig.2a.

2.2 Generative Process

The actual process to generate all triplets (w,t_a,d) which are counted in the frequency matrix $\mathrm{n}(w,t_a,d)$ is given by the graphical model depicted in Figure 2b and works as follows:

- draw a document d with probability $p(d)$;
- draw a latent topic $z \sim p(z|d)$, where $p(z|d)$ denotes the probability that a word in document d originates from topic z;
- draw the starting time $t_s \sim p(t_s|z,d)$, where $p(t_s|z,d)$ denotes the probability that the topic z starts at time t_s within the document d;
- draw a word $w \sim p(w|z)$, where $p(w|z)$ denotes the probability that a particular word w occurs within the topic z;
- draw the relative time $t_r \sim p(t_r|w,z)$, where $p(t_r|w,z)$ denotes the probability that the word w within the topic z occurs at time t_r;
- set $t_a = t_s + t_r$, which assumes that $p(t_a|t_s,t_r) = \delta(t_a - (t_s + t_r))$, that is, the probability density function $p(t_a|t_s,t_r)$ is a Dirac function. Alternatively, we could have modeled $p(t_a|t_s,t_r)$ as a noise process specifying uncertainty on the time occurrence of the word.

[1] The starting time t_s can range over different intervals, depending on hypotheses. In the experiments, we assumed that all words generated by a topic starting at time t_s occur within a document; hence t_s takes values between 1 and T_{ds}, where $T_{ds} = T_d - T_z + 1$. However, we can also assume that topics are partially observed (beginning or end are missing). In this case t_s ranges between $2 - T_z$ and T_d.

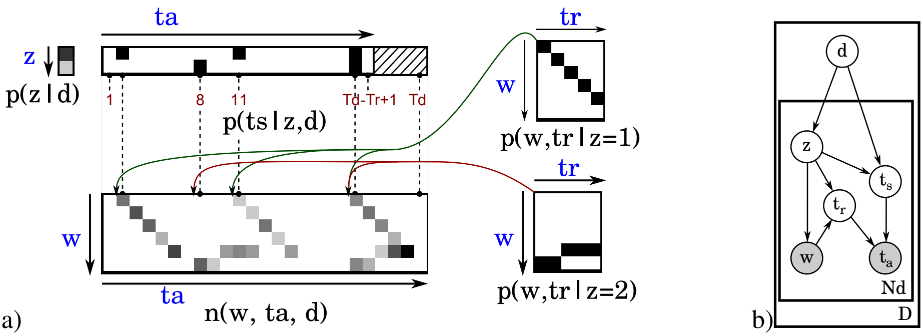


Figure 2: Generative process. a) Illustration of the document $\mathrm{n}(w,t_a,d)$ generation. Words $(w,t_a = t_s + t_r)$ are obtained by first sampling the topics and their starting times from the $p(z|d)$ and $p(t_s|z,d)$ distributions, and then sampling the word and its temporal occurrence within the topic from $p(w,t_r|z)$. b) graphical model.

The main assumption with the above model is that, given the topic, the occurrence of words within the document is independent of the topic start; that is, the occurrence of a word only depends on the topic, not on the time when a topic occurs. The joint distribution of all variables can be derived from the graphical model. However, given the deterministic relation between the three time variables ($t_a = t_s + t_r$), only two of them are actually needed to specify this distribution (for instance, we have $p(w,t_a,d,z,t_s,t_r) = p(t_r|w,t_a,d,z,t_s)p(w,t_a,d,z,t_s) = p(w,t_a,d,z,t_s)$ if $t_a = t_s + t_r$, and 0 otherwise). In the following, we will mainly use t_s and t_a for writing convenience. Accordingly, the joint distribution is given by:

$$p(w,t_a,d,z,t_s) = p(d)p(z|d)p(t_s|z,d)p(w|z)p(t_a - t_s|w,z) \tag{1}$$

2.3 Model inference

Ultimately our goal is to discover the topics and their starting times given the set of documents $\mathrm{n}(w,t_a,d)$. This is a difficult task since the topic occurrences in the documents overlap temporally, as illustrated in Fig.2a. The estimation of the model parameters Θ can be done by maximizing the log-likelihood of the observed data $\mathscr{D}$, which is obtained through marginalization over the hidden variables $Y = \{t_s, z\}$ (since $t_r = t_a - t_s$, see discussion above):

$$\mathscr{L}(\mathscr{D}|\Theta) = \sum_{d=1}^{D} \sum_{w=1}^{N_w} \sum_{t_a=1}^{T_d} n(w,t_a,d) \log \sum_{z=1}^{N_z} \sum_{t_s=1}^{T_{ds}} p(w,t_a,d,z,t_s) \tag{2}$$

The above equation can not be solved directly due to the summation terms inside the log. Thus, we employ an Expectation-Maximization (EM) approach and maximize the expectation of the complete log-likelihood instead, which is given by:

$$E[\mathscr{L}] = \sum_{d=1}^{D} \sum_{w=1}^{N_w} \sum_{t_a=1}^{T_d} \sum_{z=1}^{N_z} \sum_{t_s=1}^{T_{ds}} \mathrm{n}(w,t_a,d) p(z,t_s|w,t_a,d) \log p(w,t_a,d,z,t_s) \tag{3}$$

In the E-step, the posterior distribution of hidden variables is then calculated as:

$$p(z,t_s|w,t_a,d) = \frac{p(w,t_ad,z,t_s)}{p(w,t_a,d)} \text{ with } p(w,t_a,d) = \sum_{z=1}^{N_z} \sum_{t_s=1}^{T_{ds}} p(w,t_a,d,z,t_s) \tag{4}$$

where the joint probability is given by Eq. 1. Then, in the M-step, the model parameters (the probability tables) are updated according to (using the most convenient time variables, see end of Section 2.2):

$$p(z|d) \propto \sum_{t_s=1}^{T_{ds}} \sum_{t_r=0}^{T_z-1} \sum_{w=1}^{N_w} \mathrm{n}(w, t_s + t_r, d) p(z, t_s | w, t_s + t_r, d) \tag{5}$$

$$p(t_s|z,d) \propto \sum_{w=1}^{N_w} \sum_{t_r=0}^{T_z-1} \mathrm{n}(w, t_s + t_r, d) p(z, t_s | w, t_s + t_r, d) \tag{6}$$

$$p_w(w|z) \propto \sum_{d=1}^{D} \sum_{t_s=1}^{T_{ds}} \sum_{t_r=0}^{T_z-1} \mathrm{n}(w, t_s + t_r, d) p(z, t_s | w, t_s + t_r, d) \tag{7}$$

$$p_{t_r}(t_r|w,z) \propto \sum_{d=1}^{D} \sum_{t_s=1}^{T_{ds}} \mathrm{n}(w, t_s + t_r, d) p(z, t_s | w, t_s + t_r, d) \tag{8}$$

In practice, the EM algorithm is initialized using random values for the model parameters and stopped when the data log-likelihood increase is too small. A closer look at the above equations shows that qualitatively, in the E-step, the responsibilities of the topic occurrences in explaining the word pairs (w,t_a) are computed (where high responsibilities will be obtained for informative words, i.e. words appearing in only one topic and at a specific time), whereas the M-steps aggregates these responsibilities to infer the topic occurrences and the topic patterns. It is important to notice that thanks to the E-steps, the multiple occurrences of an activity in documents are implicitly aligned in order to learn its pattern.

Once the topics are learned, their time occurrences in any new document (represented by $p(z|d_{new})$ and $p(t_s|z,d_{new})$) can be inferred using the same EM algorithm, but using only Eq. 5 and Eq. 6 in the M-step.

3 Experiments on synthetic data

Synthetic data is used to demonstrate the strength of our model. Using a vocabulary of 10 words, we created five topics with duration ranging between 6 and 10 time steps (see Fig. 3a). Then, we created a document of 2000 time steps following the generative process described in Section 2.2, assuming equiprobable topics and 60 random occurrences per topic. One hundred time steps of this document are shown in Fig. 3b, where the intensities represents the word count (larger counts are darker), and Fig. 3g shows the corresponding starting times of three out of the five topics. As can be noticed, there is a large amount of overlap among topics. Using this “clean” document, and assuming that there are five topics with a maximum length T_z of 10, our algorithm discovers the topics and their start times perfectly.

Robustness to Noise: Two types of noise were used to test the method’s robustness. In the first case, words were added to the clean document by random sampling of the time instant t_a and the word w from a uniform distribution, as illustrated in Fig. 3d. The objective is to measure the algorithm’s performance when the ideal co-occurrences are disturbed by random word counts. The amount of noise is quantified by the ratio N_w^{noise}/N_w^{true} where, N_w^{noise} denotes the number of noise words added and N_w^{true} the number of words in the clean document. The learning performance is evaluated by measuring, under increasing noise levels, the average correlation between the learned topics $\hat{p}(t_r,w|z)$ and the true topic $p(t_r,w|z)$ (i.e. $\frac{1}{N_z}\sum_{t_r,w}\hat{p}(t_r,w|z).p(t_r,w|z)$) (See Fig. 4). We observe that the method learns the patterns effectively, as illustrated in Fig. 3c where the true topics are recovered up to the uniform noise accounting for the presence of the noisy words, and even under severe noise conditions (the correlation is almost 0.7 at a ratio of 2, i.e, when the noise words are twice as many as topic words).

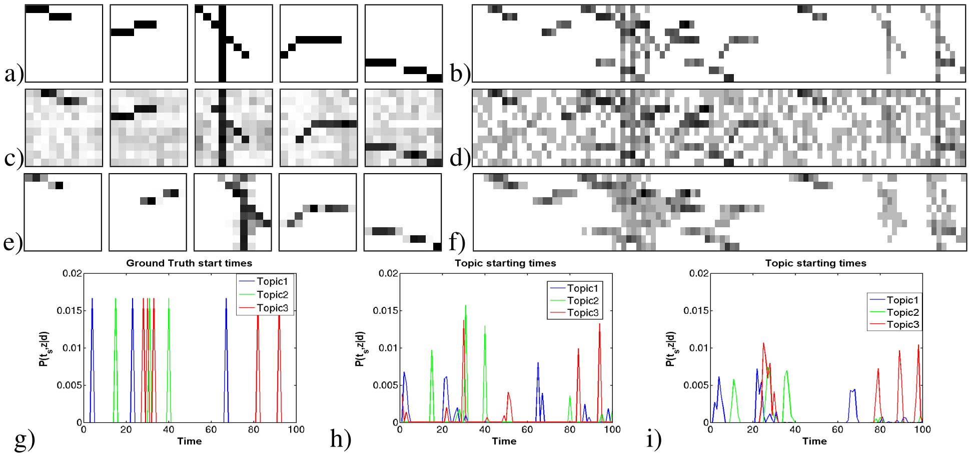


Figure 3: Synthetic experiments. (a,b,g) Clean data: (a) the five topics, (b) a segment of the generated document, (g) the topic occurrences (only 3 of them are shown for clarity). (c,d,h) experiments with uniform noise (the signal to noise ratio N_w^{noise}/N_w^{true} is 1): (c) the learnt topics, (d) the same segment as in (b) perturbed with noise, (h) the recovered topic occurrences $p(z,t_s|d)$. (e,f,i) (e) the learnt topics, (f) the noisy segment, and (i) the recovered topic occurrences, when adding a gaussian noise ($\sigma^2 = 1$) on each word time occurrence t_a.

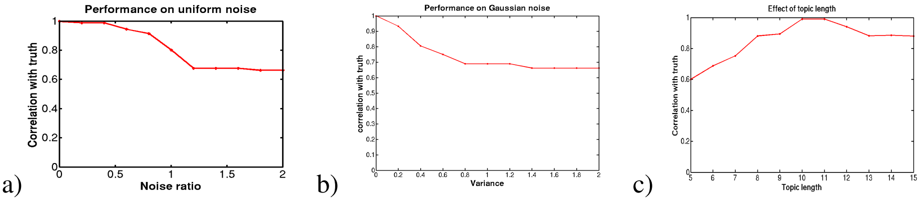


Figure 4: Average topic correlation between the estimated and the ground truth topics for different levels of (a) the uniform noise, (b) the Gaussian noise on a word time occurrence t_a, (c) Effect of varying topic length T_z from 5 to 15.

In practice, noise can also be due to variability in the temporal execution of the activity. This was simulated by adding random shifts (sampled from Gaussian noise with $\sigma \in [0,2]$) to the time occurrence t_a of each word, resulting in blurry documents (see Fig. 3f). Fig. 4b shows that our method can handle such noise. The temporal localization uncertainty is transfered partly to the topics (estimated topics are slightly perturbed versions of the true ones, see Fig. 3e) and partly to the topic occurrences (the peaks get wider in Fig. 3i).

Topic Length. The effect of varying the maximum duration T_z of a topic is summarized in the plot 4c. When T_z becomes lower than the actual topic duration, the recovered topics are truncated versions of the original ones, and the 'missing' parts are captured elsewhere, resulting in a correlation decrease. On the other hand, longer temporal windows do not really affect the learning. However, having a length much longer than the actual topic duration could result in capturing co-occurrences not originally present in the topic explaining the slow decrease in correlation.

Number of topics. Fig. 5 illustrates the algorithm's behavior when the number of topics N_z is different from the true value. When N_z is too low, we observe that each estimated topic consistently captures several true topics (e.g. first topic in Fig. 5a merges the 1st and 5th

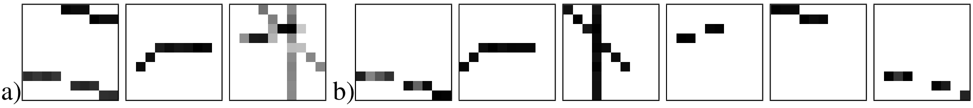


Figure 5: Estimated topics sorted by their $p(z|d)$ values when the number of topics is (a) $N_z = 3$. True topics are merged. (b) $N_z = 6$. A duplicate version of a topic with slight variation is estimated.

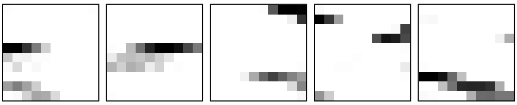


Figure 6: Five topics obtained from method [9] with clean data.

topic of Fig. 3a), while when it is too large, it usually captures variants of the same topic.

Comparison with TOS-LDA [9]. Fig. 6 shows the topics extracted from clean data by the method in [9]. This method applies the standard LDA model on documents of $N_w \times T_z$ words built from (w, t_r) pairs, where the documents consist of the temporal windows of duration T_z collected from the 'full' document. Thus, in this approach, an observed activity is represented by different sets of words depending on its relative time occurrence within these sliding windows. Or in other words, *several* topics (being time shifted versions of each other) are needed to capture the *same* activity and account for the different times at which it can occur within the window. Hence, due to the method's inherent lack of alignment ability, none of the five extracted topics truly represents one of the five patterns used to create the documents.

4 Scene activity patterns

4.1 Activity words

We also applied our pLSM model to discover temporal activity patterns from real life scenes. The flowchart is shown in Fig. 7. To apply the pLSM model on videos, we need to define the words w forming its vocabulary. One possibility would be to define some quantized low-level motion features and use these as our words. However, this would result in a redundant and unnecessarily large vocabulary. Instead, we first perform a dimensionality reduction step by extracting temporally and spatially localized activity (TSLA) patterns from the low-level features and use the occurrences of these as our words to discover sequential activity motifs (SM) in our pLSM model. To do so, we follow the approach in [12, 14] and apply a standard pLSA procedure to discover N_A dominant TSLA patterns through raw co-occurrence analysis of low-level feature words ω.

More precisely, the visual words come from the location cue (quantized into 10×10 non-overlapping cells) and the motion cue. Background subtraction is performed to identify foreground pixels, in which optical flow features are computed using the Lucas-Kanade algorithm. Foreground pixels are then categorized into either static pixels (static label) or pixels moving into a predefined direction (four labels: left, right, up, down) by thresholding the flow vectors. Thus, each word $\omega_{c,m}$ is indexed by the location c and motion m giving us $28 \times 36 \times 5 = 5040$ words.

We then apply the pLSA algorithm on a word-frequency matrix $\mathrm{n}(d_{t_a}, \omega)$ using overlapping video clips of f frames indexed by t_a as our documents, and we obtain N_A multinomial distributions $p(\omega|y)$. All topics were further automatically split using a simple connected

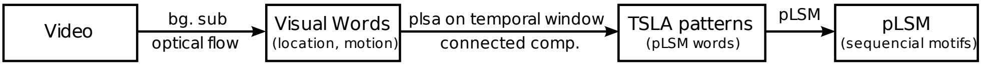


Figure 7: Flowchart for discovering sequential activity motifs in videos.

component analysis applied to their non-zero probability cells, resulting in N'_A TSLA patterns also represented by a distribution $p(\omega|y)$, $y = 1...N'_A$ (after proper normalization).

Finally, the pLSM words are defined by the TSLA patterns (i.e. $w = y$ and $N_w = N'_A$). The word counts defining the documents d are then built from the amount of presence of these TSLA patterns in the sequence of d_{t_a} documents, defined as a normalized correlation: $\mathrm{n}(d,t_a,w) = \frac{1}{\sum_{\omega \in W_y} \mathrm{n}(d_{t_a},\omega)} \sum_{\omega} \mathrm{n}(d_{t_a},\omega) p(\omega|y)$, where W_y denotes the set of non-zero probability words in the TSLA y. Note that since the correlation is conducted only on words present in W_y, it is independent of the activities going on elsewhere in the scene. This will not be the case if we use the outcome of pLSA directly as our count vector (e.g. by setting $\mathrm{n}(d,t_a,w) = p(y|d_{t_a})$).

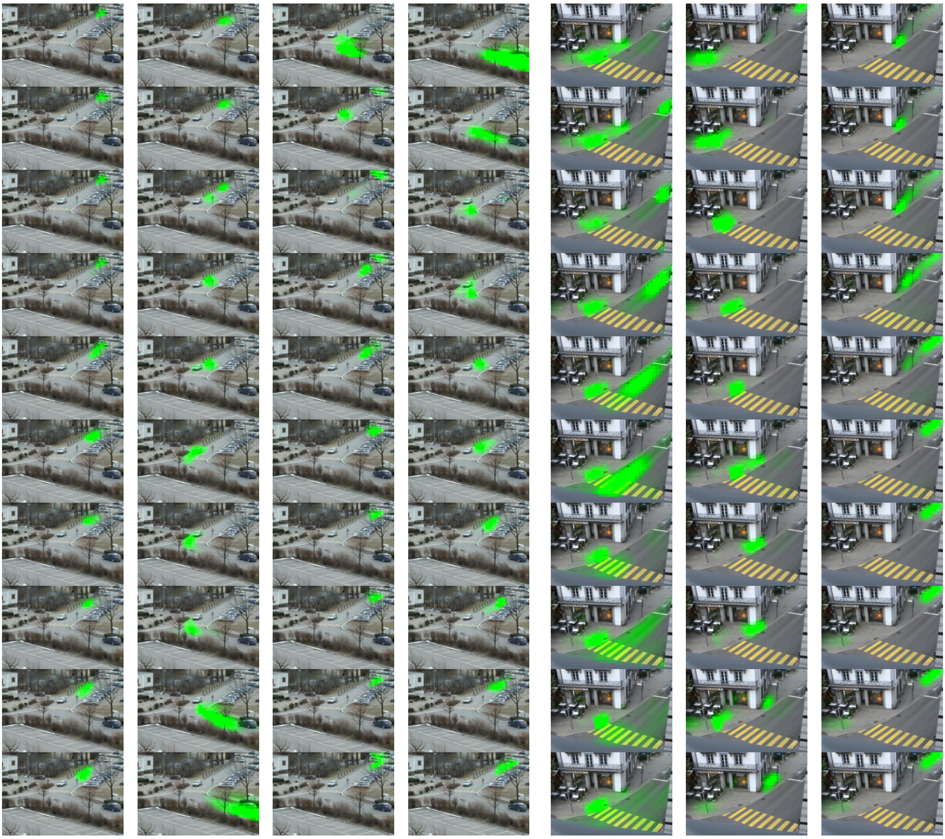
4.2 Results

Experiments were carried out on two complex scenes. The **Far Field** video contains 108 minutes of a three-road junction captured from a distance, where typical activities are moving vehicles (see Fig. 1a). As the scene is not controlled by a traffic signal, activities have large temporal variations. The **Traffic Junction** video (see Fig. 1b) is 45 minutes long and captures a portion of a busy traffic-light-controlled road junction. Activities include people walking on the pavement or waiting before crossing over the zebras, and vehicles moving in and out of the scene. Both dataset videos have a frame size of 280 × 360

In both cases, video clips of 1 second were considered to apply the first level pLSA with $N_A = 25$, which, after the connected component analysis, produced $N'_A = 27$ and $N'_A = 39$ localized TSLA patterns for the far field and traffic junction scenes, respectively. For the pLSM step, we created documents of $T_d = 150$ (2.5 minutes) and experimented with topics from $T_z = 10$ to 60 time steps. We present the topics obtained with 10 time steps (10 seconds). Given the scene complexity and the expected number of typical activities, we arbitrarily set the the number N_z of sequential motifs (SM) to 15. When increasing the number of time steps, most SMs remain the same, as observed with synthetic data. Still, some real activities that take around 20 to 25 seconds (ex. full vehicle path on Far field) are captured. In this case, a single motif captures multiple short motifs. Keeping the number of topics constant, the topics thus get redundant in the same way, as observed with synthetic data.

Figure. 8 shows the top ranking SMs from both scenes, where locations of low-level words indirectly identified at each time step are highlighted in green. The obtained SMs correspond to the dominant patterns in the scene namely, vehicle moving along the main road in both directions in the far field data. Interestingly, we see that the SMs capture the same activity occurring at different paces (e.g. Fig. 8i(b) is the same activity as in Fig. 8i(a) but done approximately twice as fast. In the Traffic Junction scene, despite the low amount of data, the motifs represent well vehicular movements and pedestrian activities, as shown in Fig. 8ii(a,b,c). Moreover, some sequential motifs capture complex temporal interactions between vehicles and pedestrians (Fig. 8ii(a)).

Event detection. We also did a quantitative evaluation of how well pLSM can be used to detect particular events. The model can estimate the most probable occurrences $p(t_s,z|d)$ of a topic z for a test document d. We can create an event detector by considering all t_s for which $p(t_s,z|d)$ is above a threshold. By varying this threshold we can control the trade-off



i(a) i(b) i(c) i(d) ii(a) ii(b) ii(c)

Figure 8: sequential motifs (SM). Each image in a column corresponds to a timestep in the pattern. The locations of the low-level words ω belonging to the words (TSLA patterns y) present in a SM at one time step are highlighted in green. (i) Top ranking SMs from far field data. (a,b) vehicles moving into the scene from top right at different speed (c,d) vehicles moving from the bottom right. (ii) Top ranking SMs for Traffic Junction video. (a) vehicle starting at red light and moving down, while pedestrians are waiting and can also cross the road after the passage of the car (b,c) pedestrians moving along the side walk.

between precision and completeness. For this event detection task, we labelled a 10 minute video clip from the far field scene, distinct from the training set, and considered 4 events depicted in Fig. 9. Note that some topics are absent of this test set (as the one in Fig. 8i(b)). To each event type, we manually associated a topic, built an event detector and varied the decision threshold to obtain precision/recall curves. Fig. 9 shows the obtained results. These results are encouraging and could be improved by post processing on $p(t_s, z|d)$. Simple non-maxima suppression can increase the precision of the detector but, in practice, it removes positive detections when cars are following each other closely and trigger a same topic at two neighboring time steps.

5 Conclusion

In this paper we proposed a novel unsupervised approach to discover dominant sequential activity motifs from video scenes. Our model infers topics which not only model the co-

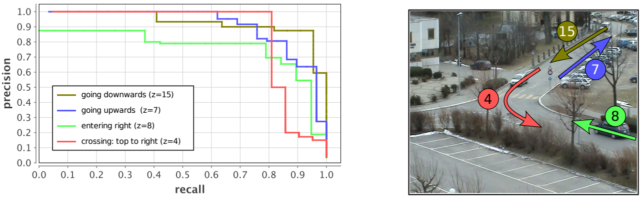


Figure 9: Interpolated Precision/Recall curves for the detection of 4 types of events mapped onto 4 topics. Evaluated on a 10 minute test video. See article's body for details.

occurrence of visual words but also the temporal order in which they appear, as well as the topic start times. Using synthetic data, various aspects of the model was analyzed and its superior performance to previous approach [9] was also demonstrated. When applied to real life data, our results are qualitatively consistent with the activities occurring in the scene. Quantitatively, performance measures also suggest the effectiveness of the method when applied to event detection task. The method's performance can further be improved in a Bayesian framework with parametric modelling. Although the method was demonstrated for activities in a video, we believe that it can have wide applications where sequential patterns need to be extracted.

6 Acknowledgements

The authors gratefully acknowledge the financial support from Swiss National Science Foundation (Project: 200020-122062) and the EU project FP7/2007-2013 VANAHEIM(248907) under which this work was carried out.

References

[1] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[2] D. M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Machine Learning Research*, (3):993–1022, 2003.

[3] Tanveer A Faruquie, Prem K Kalra, and Subhashis Banerjee. Time based activity inference using latent dirichlet allocation. In *British Machine Vision Conference*, London, UK, 2009.

[4] Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. Hidden topic markov model. *Intelligence and Statistics (AISTATS)*, March 2007.

[5] A. Hervieu, P. Bouthemy, and J.-P. Le Cadre. A statistical video content recognition method using invariant features on object trajectories. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(11):1533–1543, 2008.

[6] T. Hofmann. Unsupervised learning by probability latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[7] Timothy Hospedales, S. Gong, and Tao Xiang. A markov clustering topic model for mining behavior in video. In *ICCV*, Kyoto, Japan, 2009.

[8] Jian Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *British Machine Vision Conference*, 2008.

[9] 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*, Kyoto, Japan, 2009.

[10] Dimitrios Makris and Tim Ellis. Automatic learning of an activity-based semantic scene model. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2(1):183, 2003.

[11] C. Stauffer and E L.Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on PAMI*, 22:747–757, 2000.

[12] J. Varadarajan and J.M. Odobez. Topic models for scene analysis and abnormality detection. In *ICCV-12th International Workshop on Visual Surveillance*, 2009.

[13] X. Wang, K. Tieu, and E L. Grimson. Learning semantic scene models by trajectory analysis. *European Conference on Computer Vision*, 14(1):234–778, 2004.

[14] Xiaogang Wang, Xiaoxu Ma, and Eric L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on PAMI*, 31(3):539–555, 2009.

[15] T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *IEEE Trans. on PAMI*, 30(5):893–908, 2008.

[16] Yang Yang, Jingen Liu, and Mubarak Shah. Video scene understanding using multi-scale analysis. In *Internation Conference in Compute Vision*, Kyoto, Japan, 2009.