IJCV manuscript No.
(will be inserted by the editor)

A Sequential Topic Model for Mining Recurrent Activities from Video and Audio Data Logs.

Jagannadan Varadarajan, · Rémi Emonet · Jean-Marc Odobez

Received: date / Accepted: date

Abstract This paper introduces a novel probabilistic activity modeling approach that mines recurrent sequential patterns called *motifs* from documents given as word×time count matrices. In this model, documents are represented as a mixture of sequential activity patterns (our motifs) where the mixing weights are defined by the motif starting time occurrences. The novelties are multifold. First, unlike previous approaches where topics modeled only 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, unlike with traditional Dynamic Bayesian Networks (DBN), our model accounts for the important case where activities occur concurrently in the document (but not necessarily in synchrony), i.e. the advent of activity motifs can overlap. The learning of the motifs in these difficult situations is made possible thanks to the introduction of latent variables representing the activity starting times, enabling us to implicitly align the occurrences of the same pattern during the joint inference of the motifs and their starting times. As a third novelty, we propose a general method that favors the recovery of sparse distributions, a highly desirable property in many topic model

This work was supported by the Swiss National Science Foundation (Project: FNS-198,HAI) and from the 7th framework program of the European Union (Integrated project VANAHEIM(248907) and Network of Excellence PASCAL2). The authors gratefully thank the EU and Swiss NSF for their financial support, and all project partners for a fruitful collaboration. More information about the projects are available at the web sites www.vanaheim-project.eu and www.snf.ch.

Jagannadan Varadarajan
Idiap Research Insititute, Martigny, Switzerland
École Polytechnique Fédéral de Lausanne, Switzerland
E-mail: vjagann@idiap.ch

Rémi Emonet
Idiap Research Insititute, Martigny, Switzerland
E-mail: remonet@idiap.ch

Jean-Marc Odobez
Idiap Research Insititute, Martigny, Switzerland
École Polytechnique Fédéral de Lausanne, Switzerland
E-mail: vjagann@idiap.ch

(a) MIT data (b) Far field (c) Traffic Junction

Fig. 1: Surveillance scenes

applications, by adding simple regularization constraints on the searched distributions to the data likelihood optimization criteria. We substantiate our claims with experiments on synthetic data to demonstrate the algorithm behavior, and on three video and one audio real life datasets. We observe that using low-level motion features in the video case or Time Difference of Arrival (TDOA) features in the audio case, our algorithm is able to capture sequential patterns that implicitly represent typical trajectories of scene objects.

Keywords Unsupervised · Latent sequential patterns · Topic models · PSLA · LDA · Video surveillance

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. For instance it would be useful to define the data-driven real camera activities, to provide context for other tasks (e.g. object tracking) or to spot abnormal situations which could for instance be used to automatically select the relevant camera streams to be displayed in control rooms of public spaces monitored by hundreds of cameras.

Most activity analysis approaches are object-centered where objects are first detected, then tracked and their trajectories used for further analysis [26, 20, 37, 11]. Tracking-based approaches provide direct object-level semantic interpretation, but are sensitive to occlusion and tracking errors especially in crowded or complex scenes where multiple activities occur simultaneously, and usually require substantial computational power. Thus, as an alternative, researchers have successfully investigated algorithms relying on low-level features like optical flow that

can readily be extracted from the video stream to perform activity analysis tasks such as action recognition or abnormality detection [5, 39, 40, 19].

In visual surveillance, unsupervised methods are preferred since, due to the huge inflow of data, obtaining annotations is laborious and error prone. Such unsupervised techniques, relying on simple features like location and motion were proposed to analyze scene activities and detect abnormalities in [43, 39, 40, 19]. Among them, topic models originally proposed for text processing [12, 4] like Probabilistic Latent Semantic Analysis (PLSA) [12] or Latent Dirichlet Allocation (LDA) [4] have shown tremendous potential due to their ability to capture dominant co-occurrences in large data collections.

Topic models were first applied in vision for tasks like scene [23], object [25] and action [22] categorization. By considering quantized spatio-temporal visual features as words and short video clips as documents, they have been shown more recently to be successful at discovering scene level activities as dominant spatio-temporal co-occurrences of words. For instance, [35] used a hierarchical variant of LDA to extract atomic actions and interactions in traffic scenes, while [17] relied on hierarchical PLSA to identify abnormal activities and repetitive cycles. Activity based scene segmentation and a detailed study of various abnormality measures in this modeling context is done in [31].

Although such approaches are able to discover scene activities, the actual modeling of temporal information remains an important challenge. By relying only on the analysis of unordered word co-occurrence (due to the bag-of-words/exchangeability assumption) within a time window, most topic models fail to represent the sequential nature of activities, although activities are often temporally ordered. For example, in traffic scenes, people wait at zebra crossings 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 may be concise but not complete, as it does not allow us to distinguish it from an abnormal situation where a person crosses the road while vehicles are still moving.

In this paper, we propose an unsupervised approach based on a novel graphical topic model called *Probabilistic Latent Sequential Motifs (PLSM)*, for discovering dominant sequential activity patterns called *motifs* from sensor data logs represented by word×time counts or *temporal documents*. In this context, the main contributions of our paper are:

- a model where topics not only capture the co-occurrence of words in a temporal window, but also the *temporal order* in which the words occur within this window;
- a model that accounts for the important case where *temporal activities occur concurrently* in the document (but not necessarily in synchrony), i.e. several activities might be going on at a given time instant;
- an estimation scheme that performs *joint inference of the motifs and their starting times*, allowing us to implicitly align the occurrences of the same pattern during learning;
- a simple regularization scheme that encourages the *recovery of sparse distributions* in topic models, a higly desirable property in practice, which can be used with most topic models (e.g. PLSA, LDA).

This paper improves substantially on our work published in [30]. The improvements are mainly in the following lines: 1) The inference scheme now, uses a sparsity constraint that improves our overall results in the presence of noise as shown on the synthetic experiments; 2) A MAP formulation of the parameter estimation and a procedure to estimate the number of topics is added. 3) We have also conducted more thorough experiments on synthetic data and three real-life video datasets from state of the art papers, 4) The performance of the algorithm is quantitatively assessed and compared with other state-of-the-art models on a prediction task, 5) Lastly, we also show the generality of our model by applying it to an audio dataset.

We believe that our contribution is quite fundamental and relevant to a variety of applications where sequential motifs ought to be discovered out of time series arising from multiple activities.

The plan of the paper is as follows. In section 2, we analyse the state-of-art and compare it with our approach. Section 3 introduces our PLSM model with details, including the inference procedure. Experiments on synthetic data are first conducted in section 4 to effectively demonstrate various aspects of the model. The application of the PLSM model to the extraction of recurring activities in surveillance videos is explained in section 5, along with the presentation of the three video datasets considered for experiments. The captured PLSM motifs are shown and discussed in section 6, as well as quantitative experiments on an activity prediction task and on a comparison with ground truth labeled data. The generality of our method is further demonstrated in section 7, which present its application to audio traffic localization data captured by microphone array sensors. Finally, section 8 concludes the paper and presents some areas for future work.

2 Related Work

Our work pertains to three main issues: the modeling of activities with topic models, the discovery of temporal motifs from time series, and the learning of sparse distributions. In this Section, we briefly review the prior works conducted along these aspects and contrast them with our work.

2.1 Temporal modeling with topic models

Topic models stem from text analysis and were designed to handle large collections of documents containing unordered words. Recently, however, several approaches have been proposed to include sequential information in the modeling. This was done either to represent single word sequences [32,10], or at the high level, by modeling the dynamics of topic distributions over time [3,34,9]. For instance, [32] introduced word bigram statistics within a LDA-style model to represent topic-dependent Markov dependencies in word sequences, while in the Topic over Time method of [36], topics defined as distributions over words and time were used in a LDA model to discover topical trends over the given period of the corpus.

Many of these temporal models have been adapted for activity analysis. For instance, [13] introduced a Markov chain on scene level behaviors, but each behavior is still considered as a mixture of unordered (activity) words. More recently, [16]

used the HDP-HMM paradigm (i.e. Hierarchical Dirichlet Process, HDP, and Hidden Markov Model HMM) of [28] to identify multiple temporal topics and scene level rules. Unfortunately, for all four tested scenes only a single HMM model was discovered in practice, meaning that temporal ordering was concretely modeled at the global scene level using a set of static activity distributions, similar to what was done in [13]. Another attempt was made in [18], 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, [7] 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: firstly, only topics synchronized with respect to the cycle start can be discovered. Secondly, such a manual segmentation is time consuming and tedious. Our model addresses both these issues.

Our method is fundamentally different from all of the above approaches. The novelties are that i) the estimated patterns are not merely defined as static distributions 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; and iii) the model allows us to estimate the starting times of the activity patterns automatically.

2.2 Motifs from time series

An alternative view of activity discovery is that videos are time-series data and the various activity patterns are temporal motifs occurring in the multivariate time series. In this view, there has been some work on unsupervised activity discovery, which typically relied on HMM approaches or variants of these to perform jointly a temporal segmentation of the time series, and the learning (and identification) of the activity patterns from feature vectors. For instance, in [42] activities of individual people are clustered jointly into meeting actions using a semi-supervised layered-HMM. However, these methods assume that the entire feature vector at a given time instant corresponds to a single activity. This precludes their use in our case where multiple activities can overlap without any particular order or synchronization, resulting in a mixing of their respective features at a given time instant.

Motif discovery from time series has also been an active research area in fields as diverse as medicine, entertainment, biology, finance and weather prediction to name a few [21]. However, these methods only solve scenarios where either one or several of the following restrictions hold: there is prior knowledge about the number of patterns or the patterns themselves [15]; the data is univariate; and most importantly they assume that there is only a single pattern occurring at any time instant [27]. To the best of our knowledge, our method is one of the first attempts in discovering motifs from time series where the motifs can overlap in time.

2.3 Learning sparse distributions

One common issue in non-parametric topic models is that distributions are often loosely constrained, resulting in non-sparse process representations which are often not desirable in practice. Similar to the *sparse coding* representational scheme [41], what we seek are distributions where most of the elements in a vector are zero while few elements are significantly different from zero. For instance, in PLSA, one would like each document d to be represented by a few topics z with high weights $p(z|d)$, or each topic $p(w|z)$ to be represented by only a few words with high probability. But in practice, nothing guides the learning procedure towards such a goal. The same applies to LDA models despite the presence of priors on the multinomial $p(z|d)$ [33].

Approaches to this problem have been proposed in areas related to topic models. In Non-negative Matrix Factorization (NMF), a non-probabilistic model close to PLSA, [14] proposed to set and enforce through constrained optimization an a-priori sparsity level defined by a relationship between the L1 and L2 norm of the matrices to be learned. Very recently, [33] introduced a model that decouples the need for sparsity and the smoothing effect of the Dirichlet prior in HDP, by introducing explicit selector variables determining which terms appear in a topic. The even more complex focused topic model of [38] similarly addresses sparsity for hierarchical topic models but relies on an Indian Buffet Process to impose sparse yet flexible document topic distributions.

To address the sparsity issue, we propose an alternative approach. The main idea is to guide the learning process towards sparser (more peaky) distributions characterized by smaller entropy. We achieve this by adding a regularization constraint in the EM optimization procedure that favors lower entropy distributions by maximizing the Kullback-Leibler distance between the uniform distribution (maximum entropy) and the distribution to be learned. This results in a simple procedure that can be applied to most topic models where a sparsity constraint on the distribution is desirable.

3 Probabilistic Latent Sequential Motif Model

In this section, we first introduce notation and an overview of the model, we then describe with more details the generative process of our model, and the EM steps derived to infer the parameters of the model, including the handling of sparsity, exploitation of priors, and model selection.

3.1 Notation and model overview

Fig. 2(a) illustrates how documents are generated in our approach. Let D be the number of documents in the corpus indexed by d, 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 $t_a \in [1, \ldots, T_d]$. A document is then described by its count matrix $n(w, t_a, d)$ indicating the number of times a word w occurs at the absolute time t_a within the document. According to our model, these documents are generated from a set of N_z motifs $\{z_i\}_{i=1}^{N_z}$ represented by temporal patterns

$p(w, t_r|z)$ with a fixed maximal duration of T_z time steps (i.e. $t_r \in [0, \ldots, T_z - 1]$), where t_r denotes the relative time at which a word occurs within a topic. A topic can occur and start at any time instant $t_s \in [1, \ldots, T_{ds}]$ within the document[1]. In other words, qualitatively, documents are generated by taking the topic patterns and reproducing them in a probabilistic way (through sampling) at their starting positions within the document, as illustrated in Fig. 2(a).

3.2 Generative Process

The actual process to generate all triplets (w, t_a, d) which are counted in the matrix $n(w, t_a, d)$ is given by the graphical model depicted in Fig. 2(b) (shaded circles represent observed variables and blanc circles indicate latent variables) 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 and relative time pair $(w, t_r) \sim p(w, t_r|z)$, where $p(w, t_r|z)$ denotes the joint probability that a word w occurs at time t_r within the topic z. Note that since $p(w, t_r|z) = p(t_r|z)p(w|t_r, z)$, this draw can also be done by first sampling the relative time from $p(t_r|z)$ and then the word from $p(w|t_r, z)$, as implied by the graphical model of Fig. 2(b);
- 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.

The main assumption with the above model is that, given the motifs, the occurrence of words within the document is independent of the motif start; that is, the occurrence of a word only depends on the motif, not on the time when a topic occurs. We refer to the distribution $p(w, t_r|z)$ as *motifs* due to the temporal aspect associated to each word and to distinguish them from simple word distributions $p(w|z)$ which are used in models like PLSA/LDA.

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

$$\begin{aligned} p(w, t_a, d, z, t_s, t_r) &= p(t_r|w, t_a, d, z, t_s)p(w, t_a, d, z, t_s) \\ &= \begin{cases} p(w, t_a, d, z, t_s) & \text{if } t_r = t_a - t_s \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

[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.

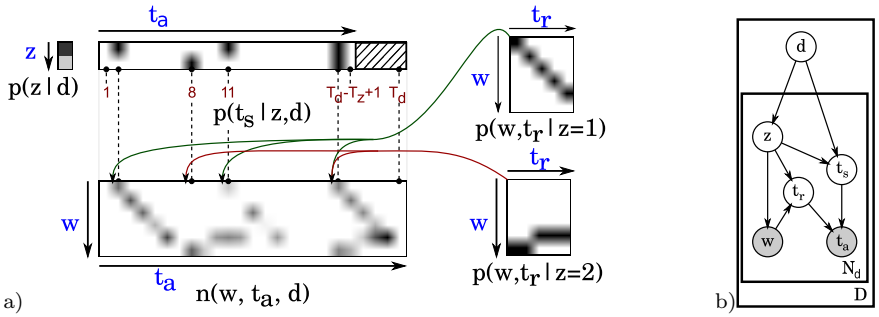
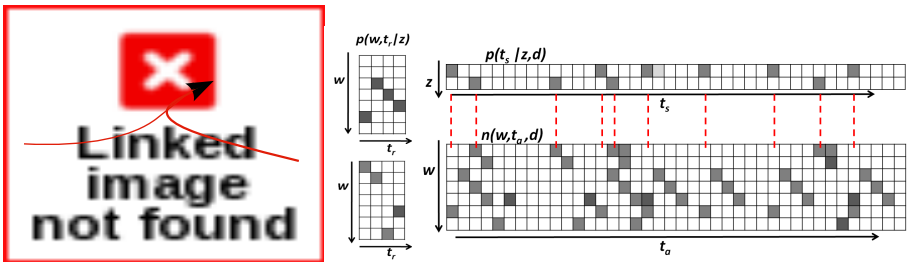


Fig. 2: Generative process. a) Illustration of the document $n(w, t_a, d)$ generation. Words $(w, t_a = t_s + t_r)$ are obtained by first sampling the topics and their starting times from the $p(z|d)$ and $p(t_s|z, d)$ distributions, and then sampling the word and its temporal occurrence within the topic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



(a) Words from two activities shown as colored blobs

(b) left: $p(w|t_r, z)$, right-top:$p(t_s|z, d)$, right-bottom:$n(w, t_a, d)$

Fig. 3: Illustration: Applying PLSM to discover activities from videos. Two activities from the scene are indicated by the red arrows. The colored blobs form our vocabulary labeled $\{w_1, ..w_7\}$. Every activity occurrence leads to a trail of observations. The count matrix $n(w, t_a, d)$ captures these observations. The latent structure is revealed through motifs $p(w, t_r|z)$ and their start times $p(t_s|z, d)$.

In the following, we will mainly use t_s and t_a. Accordingly, the joint distribution is given by:

$$p(w, t_a, d, z, t_s) = p(d)p(z|d)p(t_s|z, d)p(w, t_a - t_s|z). \tag{2}$$

To understand how PLSM applies to discovery of activities from a video, let us consider two different activities that commonly occurs in the scene above indicated by red the arrows in Fig. 3(a). *Activity A* shows a vehicle approaching from the bottom of the scene and taking a right turn. *Activity B* shows a vehicle coming from the left to right. For the sake of illustration, let us consider that our vocabulary consists of only $N_w = 7$ words, which are indicated as colored blobs with their labels $\{w_1, \ldots, w_7\}$. One may see from this word representation that when either of the activities occur, a trail of words in a particular order as $\{w_6, w_3, w_4, w_5\}$ for event A, or $\{w_1, w_2, w_7, w_5\}$ for event B can be observed. The count matrix $n(w, t_a, d)$ in Fig. 3(b-right) shows a simple case where observations from multiple occurrences of the two activities occur. However it throws little light on what

are the activities (dominant patterns) or when they occur. The count matrix also makes it clear that there are multiple events occurring at the same time without any particular synchronization, sharing the same vocabulary and often accompanied with noise. Our goal in this difficult scenario is to recover the latent structure by learning these patterns $p(w, t_r|z)$ as in Fig. 3(b-left) called motifs and their time of occurrences in $p(t_s|z, d)$ as in Fig. 3(b-right-top).

3.3 Model inference with sparsity constraints

Ultimately our goal is to discover the motifs and their starting times given the set of documents $n(w, t_a, d)$. This is a difficult task since the motif occurrences in the documents overlap temporally, as illustrated in Fig. 2(a). The estimation of the model parameters Θ, i.e., the probability distributions $p(z|d)$, $p(t_s|z, d)$, and $p(w, t_r|z)$ can be done by maximizing the log-likelihood $\mathcal{L}(\mathcal{D}|\Theta)$ of the observed data $\mathcal{D}$, which is obtained through marginalization over the hidden variables $Y = \{t_s, z\}$ (since $t_r = t_a - t_s$, as discussed at the end of the previous subsection):

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

However, as motivated in the introduction, the estimated distributions may exhibit a non-sparse structure that is not desirable in practice. In our model this is the case of $p(t_s|z, d)$: one would expect this distribution to be peaky, exhibiting high values for only a limited number of time instants t_s. To encourage this, we propose to guide the learning process towards sparser distributions characterized by smaller entropy. This could be done by adding an entropy constraint to the data likelihood. However, as this does not lead to a simple optimization, we preferred to achieve this indirectly by adding a regularization constraint in the data likelihood equation a regularization constraint to maximize the Kullback-Leibler (KL) divergence $D_{KL}(U||p(t_s|z, d))$ between the uniform distribution U (maximum entropy) and the distribution of interest. Though such an approach can be applied to any distribution of the model, we demonstrate this by applying it to $p(t_s|z, d)$. After development and removing the constant term, our constrained objective function is now given by:

$$\mathcal{L}_c(\mathcal{D}|\Theta) = \mathcal{L}(\mathcal{D}|\Theta) - \sum_{t_s,z,d} \frac{\lambda_{z,d}}{T_{ds}} \cdot \log(p(t_s|z, d)) \tag{4}$$

where $\lambda_{z,d}$ denotes a weighting coefficient balancing the contribution of the regularization compared to the data log-likelihood.

This could be done by adding an entropy constraint or a symmetric distance measure like Hellinger distance or Bhattacharyya distance to the data likelihood. But this does not lead to a simple optimization procedure. On the other hand using KL divergence helps in easily eliminating the $p(t_s|z, d)$ factors yielding a mathematically tractable form in the maximization step (see Eq. 8), which is not be possible with other distance measures. More importantly, KL divergence of the form $D_{KL}(P||Q)$ measures the error in approximating the true distribution P with a distribution Q. In our case the Uniform distribution plays the role of the

true distribution and $p(t_s|z,d)$ as an approximation to this. But by maximizing their divergence, we seek a maximum error approximation to the Uniform distribution that achieves our goal of sparsity. This also gives a clear mathematical interpretation that is in line with the theory of KL divergence.

As is often the case with mixture models, Eq. (4) can not be solved directly due to the summation terms inside the logarithm. Thus, we employ an Expectation-Maximization (EM) approach and maximize the expectation of the (regularized) complete log-likelihood instead, defined as:

$$E[\mathcal{L}] = \sum_{d=1}^{D}\sum_{w=1}^{N_w}\sum_{t_a=1}^{T_d}\sum_{z=1}^{N_z}\sum_{t_s=1}^{T_{ds}} n(w,t_a,d)p(z,t_s|w,t_a,d)\log p(w,t_a,d,z,t_s) - \sum_{t_s,z,d}\frac{\lambda_{z,d}}{T_{ds}}\cdot\log(p(t_s|z,d)) \quad (5)$$

The solution is obtained by iterating Eqs. (6–9). In the Expectation step, the posterior distribution of hidden variables is calculated as in Eq. (6) where the joint probability is given by Eq. (2). In the Maximization step the model parameters are updated by maximizing Eq. (5) along with the constraint that each of the distributions sum to one. The update expressions are given by Eqs. (7–9).

E-step:

$$p(z,t_s|w,t_a,d) = \frac{p(w,t_a,d,z,t_s)}{p(w,t_a,d)} \text{ with } p(w,t_a,d) = \sum_{z=1}^{N_z}\sum_{t_s=1}^{T_{ds}} p(w,t_a,d,z,t_s) \quad (6)$$

M-step:

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

$$p(t_s|z,d) \propto \max\left(\varepsilon, \sum_{w=1}^{N_w}\sum_{t_r=0}^{T_z-1} n(w,t_s+t_r,d)p(z,t_s|w,t_s+t_r,d) - \frac{\lambda_{z,d}}{T_{ds}}\right) \quad (8)$$

$$p(w,t_r|z) \propto \sum_{d=1}^{D}\sum_{t_s=1}^{T_{ds}} n(w,t_s+t_r,d)p(z,t_s|w,t_s+t_r,d) \quad (9)$$

Fig. 4: The EM algorithm steps.

In practice, the EM algorithm is initialized using random values for the topic distributions (see also next subsection) 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 motif 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 relative time), whereas the M-step aggregates these responsibilities to infer the motifs and their occurrences. It is important to notice that thanks to the E-step, the multiple

occurrences of an activity in documents are implicitly aligned in order to learn its pattern.

When looking at Eq. (8), we see that the effect of the additional sparsity constraint is to set to a very small constant ε the probability of terms which are lower than $\lambda_{z,d}/T_{ds}$ (before normalization), thus increasing the sparsity as desired. To set sensible values for $\lambda_{z,d}$ we used the rule of thumb $\lambda_{z,d} = \lambda\frac{n_d}{N_z}$, where n_d denotes the total number of words in the document, and λ the sparsity level. Note that when $\lambda = 1$, the correction term $\lambda_{z,d}/T_{ds}$ is, on average, of the same order of magnitude as the data likelihood – the term on the right hand side of Eq. (8) involving sums.

Inference on unseen documents. Once the motifs 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 keeping the motifs fixed and using only Eq. (7) and Eq. (8) in the M-step.

3.4 Maximum a-posterior Estimation (MAP)

In graphical models, Bayesian approaches are often preferred compared to maximum-likelihood (ML) ones, especially if there is knowledge about the model parameters. This is the case for methods like LDA that can improve over PLSA by using Dirichlet priors on the multinomial distributions. However, as was shown in [8] and [6], LDA is equivalent to PLSA when priors are uninformative or uniform, which is a common situation in practice.

The MAP estimation of parameters Θ can be formulated as follows:

$$\Theta_{\text{MAP}} = \arg\max_{\Theta}(\log P(\Theta|\mathcal{D}) = \arg\max_{\Theta}(\log P(\mathcal{D}|\Theta) + \log P(\Theta)) \tag{10}$$

where $P(\mathcal{D}|\Theta)$ is the likelihood term given by Eq. (3), and $P(\Theta)$ is the prior density over the parameter set. In practice, it is well known that using priors that are conjugate to the likelihood simplifies the inference problem. Since our data likelihood is defined as a product of multinomial distributions, we employ Dirichlet distributions as priors. A k dimensional random variable θ is said to follow a Dirichlet distribution parametrized by α if:

$$p(\theta|\alpha) \propto \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{11}$$

where, $0 \leq \theta_i \leq 1, \forall i$ and $\sum_i \theta_i = 1$. Note that $\frac{\alpha}{\|\alpha\|_1}$ represents the expected values of the parameter θ (where $\|\alpha\|_1$ is the L1 norm of α), and, when the Dirichlet is used as a prior over the parameters θ of a multinomial distribution, $\|\alpha\|$ denotes the strength of the prior, and can indeed be viewed as a count of virtual observations distributed according to $\frac{\alpha}{\|\alpha\|_1}$.

Application to the PLSM model. Our parameter set Θ comprises the multinomial parameters $p(w, t_r|z)$, $p(z|d)$, and $p(t_s|z, d)$. We don't have any a priori information about the motif occurrences $p(t_s|z, d)$ nor can we obtain an updated prior that is common to all the documents in a general scenario. Moreover, for this term, we employ the sparsity constraint rather than a smoothing prior. Thus,

we will use the MAP approach to set priors on the other multinomial parameters. Replacing in Eq. (4) the log-likelihood by the parameter log-posterior probability, the criterion to optimize simply becomes $\mathcal{L}_m(\mathcal{D}|\Theta) = \mathcal{L}_c(\mathcal{D}|\Theta) + \log P(\Theta)$, with the last term given by:

$$P(\Theta) \propto \prod_{d,z} P(z|d)^{\alpha_{z,d}-1} \prod_{z,w,t_r} P(w,t_r|z)^{\alpha_{w,t_r,z}-1}, \tag{12}$$

where $\alpha_{z,d}$ and $\alpha_{w,t_r,z}$ denote the Dirichlet parameters governing the prior distributions of $P(z|d)$ and $P(w,t_r|z)$ respectively. As before, $\mathcal{L}_m$ can be conveniently optimized using an EM algorithm, which leads to the same update expression as in Fig. 4, except that Eq. (7) and Eq. (9) need to be modified to account for the prior.

$$p_{\text{MAP}}(z|d) \propto (\alpha_{z,d} - 1) + \sum_{t_s=1}^{T_{ds}} \sum_{t_r=0}^{T_z-1} \sum_{w=1}^{N_w} n(w, t_s + t_r, d) p(z, t_s | w, t_s + t_r, d) \tag{13}$$

$$p_{\text{MAP}}(w,t_r|z) \propto (\alpha_{w,t_r,z} - 1) + \sum_{d=1}^{D} \sum_{t_s=1}^{T_{ds}} n(w, t_s + t_r, d) p(z, t_s | w, t_s + t_r, d) \tag{14}$$

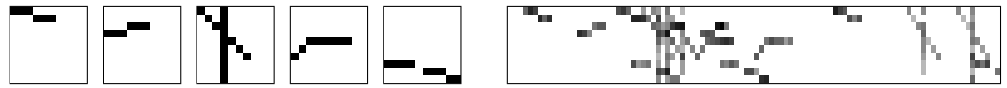
3.5 Model Selection

In unsupervised learning methods that are akin to clustering, the number of clusters is an important parameter to be determined. In our problem, this issue translates into identifying an appropriate number of motifs. Usually in real-life scenarios, we have some rough a-priori knowledge of the number of motifs. This is the case, for instance, in our video activity analysis scenarios, where this number qualitatively depends on the scene complexity, the types of features (observations), and the duration of the sought motifs. Still, being able to adapt the selected number of motifs as a function of the actual data is desirable.

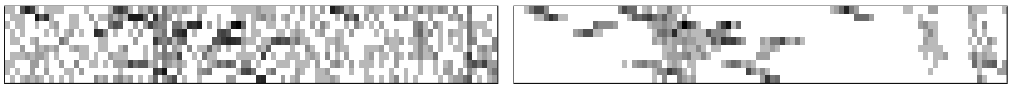
There are several methods that can be used for model selection in unsupervised settings. They include testing on held-out data [2], the Bayesian Information Criterion (BIC) [24], and more sophisticated non-parametric approaches like Hierarchical Dirichlet Processes [28]. In this work, we use the BIC measure, which penalyzes the training data likelihood based on the number of parameters and data points. The BIC measure of a model M is calculated as:

$$BIC(M) = -2\mathcal{L}(\mathcal{D}|\Theta) + \lambda_{bic} N_p^M \log(n) \tag{15}$$

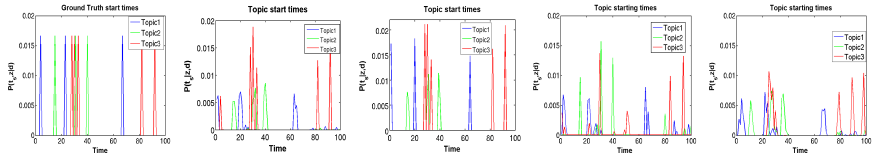
where, $\mathcal{L}$ is the likelihood of the model and is given by Eq. (3), N_p^M denotes the number of parameters of model M, n is the number of data points, and λ_{bic} is a coefficient that controls the influence of the penalty. This criterion seeks models that find a compromise between likelihood fitting and model complexity. In practice, we conduct optimization for models with different number of motifs according to previous subsections, and finally keep the model with the minimum BIC measure.



(a) Motifs to generate documents (b) Sample of a generated document



(c) Document with uniform noise (d) Document with location noise



(e) True occurrences (f) Clean data, $\lambda = 0.0$ (g) Clean data, $\lambda = 0.5$ (h) Uniform noise, $\lambda = 0.5$ (i) Location noise, $\lambda = 0.5$

Fig. 5: Synthetic experiments. (a) The five motifs, (b) A segment of a generated document, (c,d) The same segment perturbed with: (c) Uniform noise ($\sigma_{snr} = 1$), (d) Gaussian noise ($\sigma = 1$) added to each word time occurrence t_a. (e) the true motif occurrences (only 3 of them are shown for clarity) in the document segment shown in (b). (f–i) the recovered topic occurrences $p(t_s|z, d)$ when using as input (f) the clean document (cf b) and no sparsity constraint $\lambda = 0$ (g) or with sparsity constraint $\lambda = 0.5$; (h) the noisy document (c) and $\lambda = 0.5$ (i) the noisy document (d) and $\lambda = 0.5$.

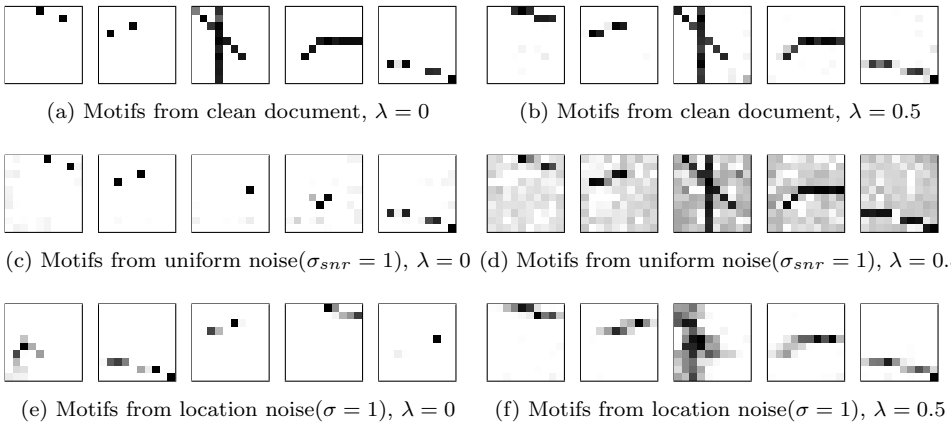
4 Experiments on synthetic data

In order to investigate and validate various aspects and strengths of the model we first conducted experiments using synthetic data.

4.1 Data and experimental protocol

Data synthesis. Using a vocabulary of 10 words, we created five motifs with duration ranging between 6 and 10 time steps (see Fig. 5(a)). Then, for each experimental condition (e.g. a noise type and noise level), we synthesized 10 documents of 2000 time steps following the generative process described in section 3.2, assuming equi-probable topics and 60 random occurrences per motif. One hundred time steps of one document are shown in Fig. 5(b), where the intensities represents the word count (larger counts are darker). In Fig. 5(e) corresponding starting times of the first three motifs out of the five motifs are shown for the sake of clarity. Note that there is a large amount of overlap between motifs.

Adding noise. Two types of noise were used to test the method's robustness. In the first case, words were added to the clean documents by randomly sampling the time instant t_a and the word w from a uniform distribution, as illustrated in Fig. 5(c). Here, the objective is to measure the algorithm's performance when the



(a) Motifs from clean document, $\lambda = 0$ (b) Motifs from clean document, $\lambda = 0.5$

(c) Motifs from uniform noise($\sigma_{snr} = 1$), $\lambda = 0$ (d) Motifs from uniform noise($\sigma_{snr} = 1$), $\lambda = 0.5$

(e) Motifs from location noise($\sigma = 1$), $\lambda = 0$ (f) Motifs from location noise($\sigma = 1$), $\lambda = 0.5$

Fig. 6: Synthetic experiments. Recovered motifs without (a,c,e) and with (b,d,f) sparsity constraints $\lambda = 0.5$ under different noise conditions. (a,b) from clean data; (c,d) from documents perturbed with random noise words, $\sigma_{snr} = 1$, cf Fig. 5(c); (e,f) from documents perturbed with Gaussian noise on location $\sigma = 1$, cf Fig. 5(d).

ideal co-occurrences are disturbed by random word counts. The amount of noise is quantified by the ratio $\sigma_{snr} = N_w^{noise}/N_w^{true}$ where, N_w^{noise} denotes the number of noise words added and N_w^{true} is the number of words in the clean document. In practice, noise can also be due to variability in the temporal execution of the activity. Thus, in the second case, a 'location noise' 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, as shown in Fig. 5(d).

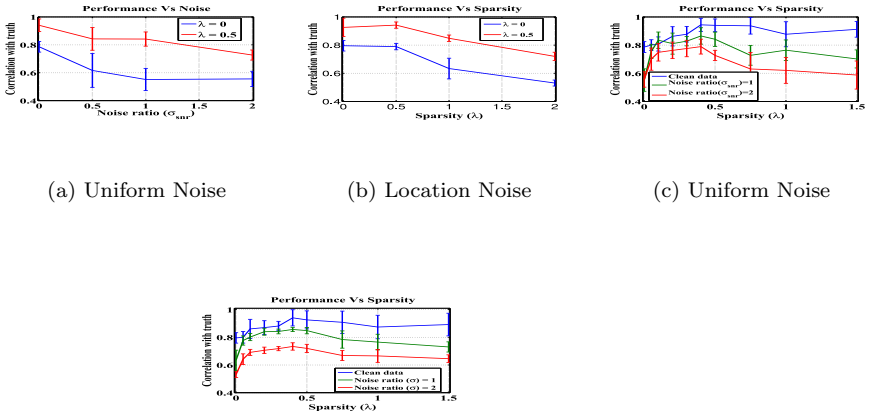
Model parameterization. As we do not assume any prior on the parameter model, we did not use the MAP approach in these experiments, and optimized the penalized likelihood of Eq. (4). For each document, 10 different random initializations were tried and the model maximizing the objective criterion was kept as the result.

Performance measure. The learning performance is evaluated by measuring the normalized cross correlation [2] Averages and corresponding error-bars computed from the results obtained on the 10 generated documents are reported.

4.2 Results

Results on clean data. Figs 6(a) and 6(b) illustrate the recovered motifs with and without the sparsity constraint. As can be seen, without sparsity, two of the obtained motifs are not well recovered. This can be explained as follows. Consider the first of the five motifs. Samples of this motif starting at a given instant t_s in the document can be equivalently obtained by sampling words from the learned

[2] The correspondence between the ground truth topics and the estimated ones is made by optimizing the normalized cross-correlation measure between the learned motifs $\hat{p}(t_r, w|z)$ and the true motifs $p(t_r, w|z)$.



(a) Uniform Noise (b) Location Noise (c) Uniform Noise

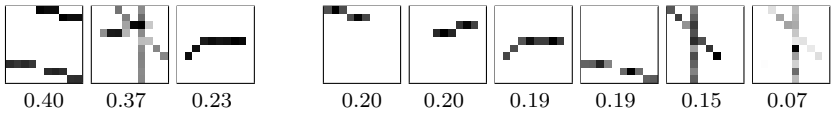
(d) Location Noise

Fig. 7: (a,b) Average motif correlation between the estimated and the ground truth motifs for different levels of (a) Uniform noise, (b) Location noise. (c,d) Average motif correlation between the estimated and the ground truth motifs for different sparsity weight λ and for different levels of (c) Uniform noise, (d) Gaussian noise on a word time occurrence t_a (Location Noise).

motif Fig. 6(a) and sampling the starting time from three consecutive t_s values with probabilities less than one. This can be visualized in Fig. 5(f), where the peaks in the blue curve $p(t_s|z = 1, d)$ are three times wider and lower than in the ground truth. When using the sparsity constraint, the motifs are well recovered, and the starting time occurrences better estimated, as seen in Fig. 6(b).

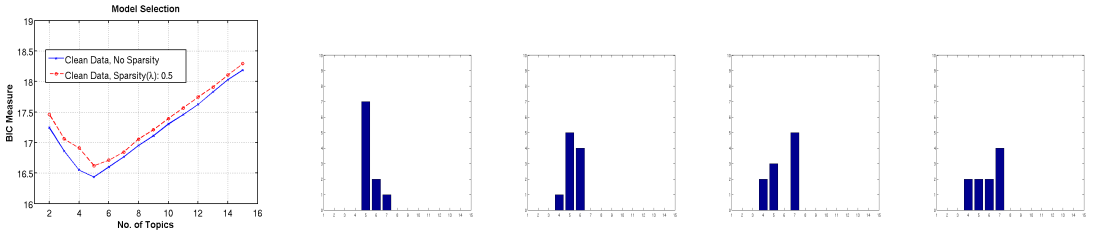
Robustness to noise. Fig. 6(c) and 6(e) illustrate the recovered motifs under noise, without a sparsity constraint. We can clearly observe that the motifs are not well recovered (e.g. the third motif is completely missed). With sparsity, Fig. 6(d) and 6(f), motifs are better recovered, but reflect the presence of the generated noise, i.e. the addition of uniform noise in the motifs in the first case, and the temporal blurring of the motifs in the second case. The curves in Fig. 7(a) and 7(b) show the degradation of the learning as a function of the noise level.

Effect of sparsity. We also analyzed the performance of the model by varying the weight of the sparsity constraint for different noise levels and noise types. Fig. 7(a) and 7(b) show that the model is able to handle quite a large amount of noise in both cases, and that the sparsity approach always provides better results. While the best results without the constraint gives only a correlation of 0.8, we achieve a much better performance (approximately 0.95) with sparsity. In Fig. 7(c) and 7(d), we see the performance of the method for various values of the sparsity weight λ and for varying noise levels. We notice that as the weight for sparsity increases, the performance shoots up. However, an increase of the sparsity weight beyond 0.5 often leads to degraded and sometimes unstable performance. Finally also note,



(a) Three motifs (b) Six motifs

Fig. 8: Estimated motifs sorted by their $p(z|d)$ values (given below each topic) when the number of motifs is (a) $N_z = 3$. True motifs are merged. (b) $N_z = 6$. A duplicate version of a motif with slight variation is estimated.



(a) BIC measure (b) clean, $\lambda = 0$ (c) clean, $\lambda = 0.5$ (d) noise, $\lambda = 0$ (e) noise, $\lambda = 0.5$

Fig. 9: (a) Example of a BIC measure on a synthetic document. (b-d) Selection histograms (number of times a topic is selected) using BIC on 10 documents, with the following conditions: (b) clean, $\lambda = 0$ (c) clean, $\lambda = 0.5$ (d) Noise ($\sigma_{snr} = 1$), $\lambda = 0$, (e) Noise ($\sigma_{snr} = 1$), $\lambda = 0.5$.

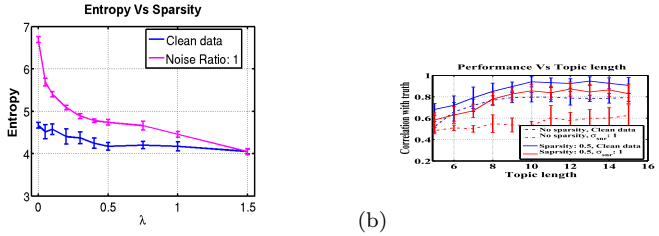
as illustrated by Fig. 10(a), that the increase of the sparsity weight λ leads to a lowering of the entropy of $p(t_s|z, d)$, as desired.

We conclude from these results that we obtain a marked improvement in recovering the motifs from both clean and noisy documents when sparsity constraint is used.

Number of topics and model selection. We first studied the qualitative effect of changing N_z, the number of motifs. As illustrated in Fig. 8. When N_z is lower than the true number, we observe that each estimated motif consistently captures several true motifs. For instance, the first motif in Fig. 8(a) merges the 1st and 5th motif of Fig. 5(a). When the number of topics is larger than the true value, like $N_z = 6$ in the example, we see that a variant of one motif is captured, but with lower probability. We observe the same phenomenon as we further increase the number of motifs.

We also tested our model selection approach based on the BIC criteria, as explained in section 3.5. To set λ_{bic}, we generated five extra clean documents and used them to select an appropriate value of this parameter. Then, the same value was used to perform tests on other clean or noisy documents. Fig. 9(a) displays the BIC values obtained for a clean document by varying the number of motifs from 2 to 15. As can be seen, the criteria reaches its minimum for 5 motifs. Histograms in Fig. 9(b-d) show the number of selected topics for a set of documents. Although not perfect, the results show that the method is able to retrieve an appropriate number of topics, and that in the presence of noise, the number of found motifs is usually larger to explain the presence of the additional noise in the data.

Motif length. The effect of varying the maximum duration T_z of a motif and in



(a) (b)

Fig. 10: (a) Average entropy of $p(ts|z,d)$ as a function of sparsity λ. (b) Effect of varying motif length T_z from 5 to 15, for two levels of uniform noise.

Fig. 11: Five motifs obtained from the method [18] with clean data.

the presence of noise is summarized in Fig. 10(b). When T_z becomes lower than the actual motif duration, the recovered motifs are truncated versions of the original ones, and the "missing" parts are captured elsewhere, resulting in a decrease in correlation. On the other hand, longer temporal windows do not really affect the learning, even under noisy conditions. However, the performance under clean and noisy conditions are significantly worse with no sparsity constraint.

Comparison with TOS-LDA [18]. Fig. 11 shows the motifs extracted from clean data by the method in [18]. 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* motifs (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 motifs truly represents one of the five patterns used to create the documents.

5 Application to video scene activity analysis

Our objective is to identify recurring activities in video scenes from long term data automatically. In this section, we explain how we can use the PLSM model for this purpose, and describe the video preprocessing used to define the words and temporal documents required by the PLSM model. We then present the datasets used for experiments and finally show three different ways of representing the learned motifs.

5.1 Activity word and temporal document construction

To apply the PLSM model to videos, we need to specify its inputs: the words w forming its vocabulary and that define the semantic space of the learned motifs,

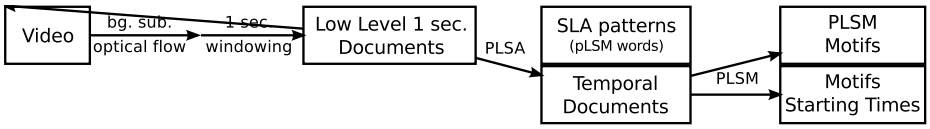


Fig. 12: Flowchart for discovering sequential activity motifs in videos. Quantized low-level features are used to build 1 second bag-of-words documents, from which Spatially Localized Activity patterns (SLA) are learned and further used to build the temporal documents used as input to PLSM.

and the corresponding temporal documents. One possibility would be to define quantized low-level motion features and use these as our words. However, this would result in a redundant and unnecessarily large vocabulary. We thus propose to first perform a dimensionality reduction step by extracting spatially localized activity (SLA) patterns from the low-level features and use the occurrences of these as our words to discover sequential activity motifs using the PLSM model. To do so, we use the approach in [35,31] and apply a standard PLSA procedure to discover N_A dominant SLA patterns through raw co-occurrence analysis of low-level visual words w^{ll}. The work flow of this process is shown in Fig. 12, and explained below.

Low-level words w^{ll}. The visual words come from the location cue (quantized into 2×2 non-overlapping cells) and motion cue. First, background subtraction is performed to identify foreground pixels, in which optical flow features are computed using Lucas-Kanade algorithm [29]. The foreground pixels are then categorized into either static pixels (static label) or pixels moving into one of the eight cardinal directions by thresholding the flow vectors. Thus, each low-level word $w^{ll}_{c,m}$ is implicitly indexed by its location c and motion label m. Note that the static label will be extremely useful for capturing waiting activities, which contrasts with previous works [35].

Low-level SLA patterns z^{ll}. We apply the PLSA algorithm on a document-word frequency matrix $n(d_{t_a}, w^{ll})$ obtained by counting for the document d_{t_a} the low-level words appearing in N_f frames within a time interval of one second centered on time t_a. The result is a set of N_A SLA patterns characterized by their multinomial distributions $p(w^{ll}|z^{ll})$, and the probabilities $p(z^{ll}|d_{t_a})$ providing the topic distribution for each document. While PLSA captures dominant low-level word co-occurrences, it can also be viewed as a data reduction process since it provides a much more concise way of representing the video underlying activities at a given instant t_a, using only N_A topics, a number much smaller than the low-level vocabulary size. In practice, we observed that between 50 and 100 SLA topics/patterns are sufficient to provide an accurate description of the scene content, and used $N_A = 75$.

We can visualize the result of this step by superimposing the distributions $p(w^{ll}|z^{ll})$ over the image, indicating the locations where they have high probabilities. This is illustrated in Fig. 13, which shows representative SLA patterns obtained from each of the three video scenes described below, with their loca-

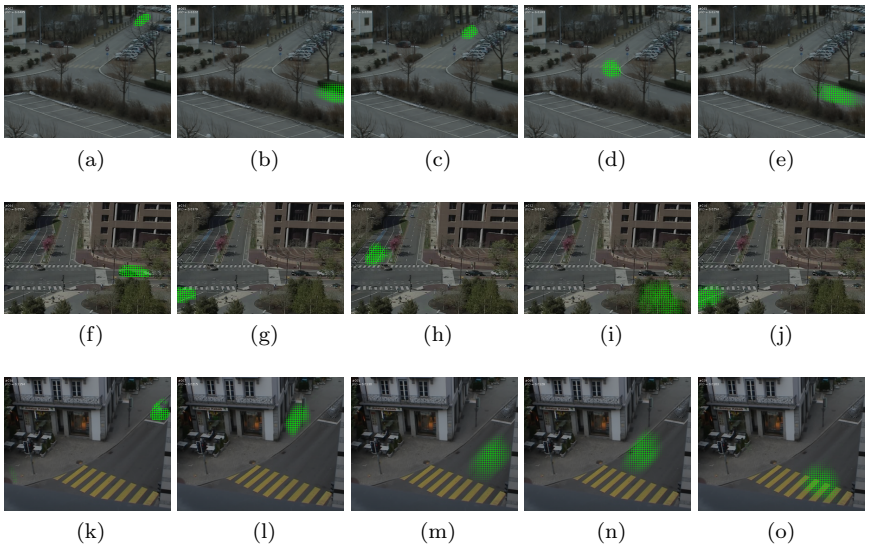


Fig. 13: Representative SLA patterns obtained by applying PLSA on (a–e) far-field data, (f–j) MIT data and (h–l) Traffic junction data.

tions highlighted in green[3]. Clearly, the SLA patterns represent spatially localized activities in the scene.

Building PLSM temporal documents. In our approach, we define the PLSM words as being the SLA patterns (i.e. we have $w \leftrightarrow z^{ll}$ and $N_w = N_A$). Thus, to build the temporal documents d for PLSM, we need to define our word count matrix $n(d, t_a, w)$ characterizing the amount of presence of the SLA patterns z^{ll} in the associated low-level document at this instant t_a, i.e. d_{t_a}. To do so, we exploit two types of information: the overall amount of activity in the scene at time t_a, and how this activity is distributed amongst the SLA patterns. The word counts were therefore simply defined as:

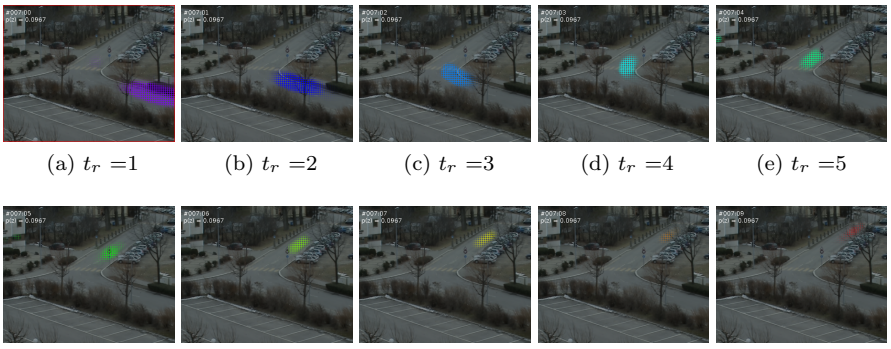
$$n(d, t_a, w) = n(d_{t_a}) p(z^{ll} | d_{t_a}) \tag{16}$$

where $n(d_{t_a})$ denotes the number of low-level words observed at a given time instant (i.e. within the 1 second interval used to build the d_{t_a} document). We set T_d to 120, and thus each temporal document is created from video clips of 2 minutes duration.

5.2 Video datasets

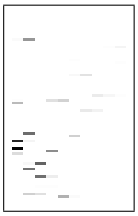
Experiments were carried out on three complex scenes with different activity contents. The **MIT** scene [35] is a two-lane and four-road junction captured from a

[3] Note that the topic distributions contain more information than the location probability: for each location, we know what types of motion are present as well. This explains the location overlap between several topics, e.g. between those of Fig. 13b and Fig. 13e, which have different dominant motion directions.

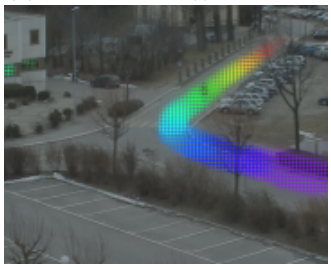


(a) t_r =1 (b) t_r =2 (c) t_r =3 (d) t_r =4 (e) t_r =5

(f) t_r =6 (g) t_r =7 (h) t_r =8 (i) t_r =9 (j) t_r =10



(k) $p(w, t_r|z)$



(l) collapsed t_r =1:10

Fig. 14: Three different representations of a PLSM motif. (k) Motif probability matrix. The x axis denotes t_r, and the y axis the words. (a-j) For each time step t_r, weighted overlay on the scene image of the locations associated to each word (i.e. the SLA patterns). (l) All time steps collapsed into one image color-coded according to the rainbow scheme, (Violet for $t_r = 1$ to Red for $t_r = T_z$).

distance, where there are complex interactions among vehicles arriving from different directions, and few pedestrians crossing the road (see Fig. 1a). This has a duration of 90 minutes, recorded at 30 frames-per-second (fps), and a resolution of 480 × 756 which was down-sampled to half its size. The **Far-field** scene [30] depicts a three-road junction captured from a distance, where typical activities are moving vehicles (see Fig. 1b). As the scene is not controlled by a traffic signal, activities occur at random. The video duration is 108 minutes, recorded at 25 fps and a 280×360 frame resolution. The **Traffic junction** [31] (see Fig. 1c) captures a portion of a busy traffic-light-controlled road junction. In addition to vehicles moving in and out of the scene, activities in this scene also include people walking on the pavement or waiting before walking across the pedestrian crossing. The video, recorded at 25 fps and a 280 × 360 frame resolution, has a duration of 44 minutes.

5.3 Motif representation

Before looking at the results obtained from the datasets, we explain how learned motifs are represented visually. In Fig. 14, we provide three different ways of representing a recovered motif of $T_z = 10$ time steps (seconds) duration obtained from PLSM. By definition, a PLSM motif is a distribution $p(w, t_r|z)$ over $w \times t_r$ space. Thus the direct depiction of the motif is that of the $p(w, t_r|z)$ matrix as given in Fig. 14(k). This shows that the distribution is relatively sparse, that words often occur at several consecutive time steps, and that several words co-occur at each time step. However, this does not provide much intuition about the activities captured by the motif. The second way of representing the motif is to back-project on the scene image and for each time step t_r, the locations associated with the words (the SLA patterns) probable at this time step, similar to the illustration of the SLA patterns in Fig. 13. This is illustrated in Fig. 14(a–j). This provides a good representation of the motif, but is space consuming. An even more realistic representation giving a true grasp of the motifs is provided by rendering them as animated gifs. This is what we provided in the additional material on the website `http://www.idiap.ch/~vjagann/plsm.html`.

Due to media and space limitations, we use here an alternative version of these representations that collapses all time step images into a single image using a color-coded scheme, as shown in Fig. 14(l). Note that the color at a given location is the one of the largest time step t_r for which the location probability is non zero. Hence, the representation may hide some local activities due to the collapsing effect. However, in the large majority of cases, the representation provides good intuition of the learned activities.

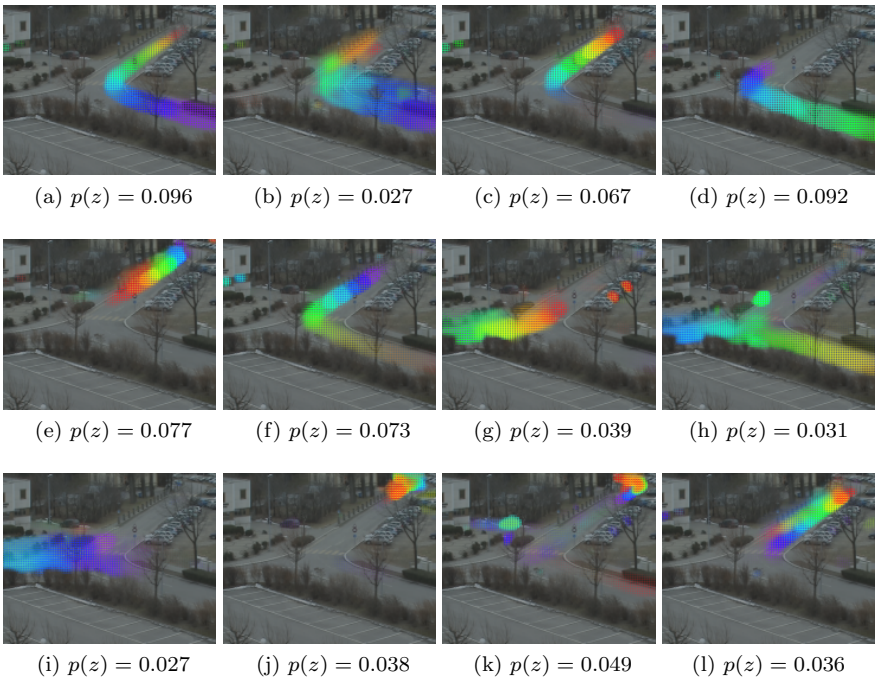
6 Video Scene Analysis Results

In this section, complementary details about the algorithm implementation are provided. Then, recovered motifs on the three datasets are shown and commented on. We then report the results of quantitative experiments on a counting task and on a prediction task to further validate our approach.

6.1 Experimental details

For the low-level processing, 1 second intervals were used to build the low-level documents and then the PLSM temporal document. To reduce the computational cost, optical flow features were estimated and collected in only $N_f = 5$ frames of these intervals. To favor the occurrence of the word probability mass at the start of the estimated motifs, we relied on the MAP framework and defined Dirichlet prior parameters for the motifs[4] as $\alpha_{w,t_r,z} = \tau \cdot \frac{1}{N_z} \cdot f(t_r)$, where f denotes a normalized (i.e. the values of $f(t_r)$ sums to 1) decreasing ramp function as $f(t_r) \propto (T_z - t_r) + c$, T_z is the motif duration and c is a constant term. In other words, we did not impose any prior on the word occurrence probability, only on the time when they can occur. The strength of the prior is given by the term τ and

[4] Note that we did not set any prior on the topic occurrences within the document, i.e. we set $\alpha_{z,d} = 0$.



(a) $p(z) = 0.096$ (b) $p(z) = 0.027$ (c) $p(z) = 0.067$ (d) $p(z) = 0.092$

(e) $p(z) = 0.077$ (f) $p(z) = 0.073$ (g) $p(z) = 0.039$ (h) $p(z) = 0.031$

(i) $p(z) = 0.027$ (j) $p(z) = 0.038$ (k) $p(z) = 0.049$ (l) $p(z) = 0.036$

Fig. 15: Far-field data. Twelve representative motifs of 10s duration, out of 20. The method is able to capture the different vehicular trajectory segments. Best viewed in color. Please see the page at [1] to view animated gif versions of the motifs.

was defined as a small fraction (we used 0.1) of the average number of observations in the training data for each of the $N_z \cdot N_w \cdot T_z$ motif bins. In practice, the prior plays a role when randomly drawing the motifs at initialization, where they are generated from the prior, and during the first EM iterations. After, given the (low) level of the τ value and the concentration of the real observations on a few motif bins (see an estimated topic in Fig. 14), its influence becomes negligible.

6.2 PLSM motifs and activities

We first searched for motifs of maximum 10 seconds duration, i.e. $T_z = 10$. Note that 10 seconds already captures relatively long activities, especially when dealing with vehicles. At the end of this Section, we also show results when looking for 20 second motifs.

The number of 10s motifs selected automatically using the BIC criteria were 20, 26 and 16 topics for the Far-field, MIT, and Traffic junction datasets respectively. A selection of the top-ranking representative motifs are shown in Fig. 15, Fig. 16,

and Fig. 17 using the collapsed color representation (cf Fig. 14), along with their probability $p(z)$ in explaining the training data.[5] Below we comment on the results.

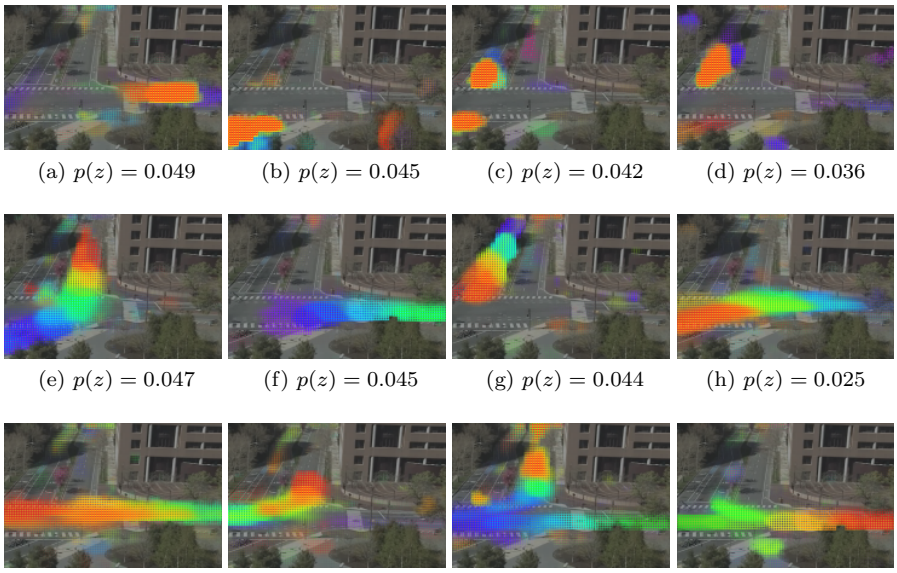
Far-field data. The analysis of the motifs show the ability of the method to capture the dominant vehicle activities and their variations due to differences of trajectory, duration, and vehicle type, despite the presence of trees at several places that perturb the estimation of the optical flow. For instance, Fig. 15(a–c) correspond to vehicles moving towards the top right of the image, and Fig. 15(d–f) to vehicles moving from the top right. Fig. 15(g) corresponds to vehicles moving from left of the scene to the top right, Fig. 15(h) to vehicles moving from left to bottom and Fig. 15(i) to movement towards the left. Some of the motifs capture almost the full presence of a vehicle in the scene (e.g Fig. 15(a,b,f,h)) which happens when vehicles move fast enough so that the duration of their appearance is close to 10s. Otherwise, when vehicles move more slowly, the model has to split their trajectory into different sub activities (generally two). Interestingly enough, we often observe that the model automatically captures segments that are common to multiple trajectories. This variability due to speed is also illustrated by the activity captured by the motif in Fig. 15(e) which is much slower than that of Fig. 15(f) since it only crosses around half the distance of the motif in Fig. 15(f), for the same duration. Motifs in Fig. 15(j,k) represent the activities of vehicles moving in and out of the scene at the top of the scene. Since this location is far from the camera, and vehicles in both directions have to slow down due to a bump in the road, their apparent motion in the image is very slow and all the words are concentrated over a small region for the entire motif duration. Finally, the motif in Fig. 15(l) represents the activity of two vehicles passing each other on the top part of the road.

MIT data. This dataset is quite complex, with multifarious activities occurring concurrently and being only partially constrained by the traffic light. Even in this case, our method extracted meaningful activities corresponding to the different phases of the traffic signal cycle, as shown in Fig. 16. Briefly speaking, one finds two main activity types: waiting activities, shown in Fig. 16(a-d)[6], and dynamic activities as shown in Fig. 16(e-l) of vehicles moving from one side of the junction to the other after the lights change to green. Note that waiting activities were not captured in previous works like [35], are identified here thanks to the use of background subtraction and of static words.

Traffic junction data. Despite the small amount of data (44min) and complex interactions between the objects of the scene, the method is able to discover the dominant activities as shown in Fig. 17. These are for instance car dynamical activities, which usually last around 5 seconds only which explains the absence of the whole color range in Fig. 17(a-c). Note that while Fig. 17(a) corresponds to cars going straight, Fig. 17(b) shows cars coming from the top right and turning to their right at the bottom. Waiting activities are also captured, as illustrated in the motif of Fig. 17(d), which displays vehicles waiting for the signal. Interestingly, another set of motifs capture pedestrian activities, despite the fact that they are less

[5] In [1], an exhaustive set of results are provided with motifs rendered in animated-GIF.

[6] Waiting activities are characterized by the same word(s) repeated over time in the motif. Thus the successive time color-coded images overwrite the previous ones in the collapsed representation as explained in Section. 5.3, leaving visible only the last (orange, red) time instant.



(a) $p(z) = 0.049$ (b) $p(z) = 0.045$ (c) $p(z) = 0.042$ (d) $p(z) = 0.036$

(e) $p(z) = 0.047$ (f) $p(z) = 0.045$ (g) $p(z) = 0.044$ (h) $p(z) = 0.025$

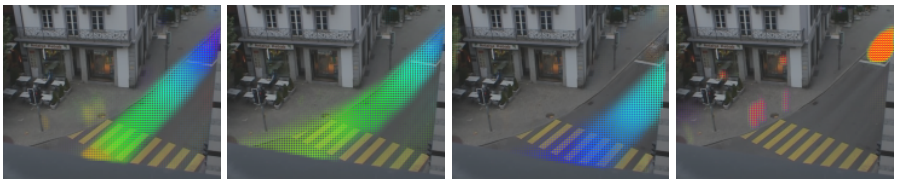
(i) $p(z) = 0.040$ (j) $p(z) = 0.039$ (k) $p(z) = 0.035$ (l) $p(z) = 0.040$

Fig. 16: MIT data. Representative motifs of 10s duration out of 26. (a–d) Activities due to waiting objects. (e–l) Activities due to motion. Best viewed in color. Please see the page at [1] to view animated GIF versions of the motifs.

constrained and have more variability in localization, size and shape, timing and dynamics. This comprises people moving on the sidewalk (Fig. 17(e,f)), but also pedestrians crossing the road on the zebra crossing as in motifs from Fig. 17(g,h).

Topic Length. We also experimented with longer motif duration T_z. For instance Fig. 18 shows motifs of 20 second duration from all the three datasets. Since longer motifs capture more activities, the BIC measure selected only 16, 16 and 14 motifs for the Far-field, MIT, and Traffic junction data respectively. Broadly speaking, when one extends the motif maximal length beyond the actual duration of a scene activity, the same motif is estimated, as already observed with synthetic data[7]. This is typically the case with the short vehicle motifs in the MIT (Fig. 16(f,l)) or Traffic junction (Fig. 17(a-c)) datasets. Still, as activities can often be described with different time granularities, variations or other motifs may appear. For instance, as the travel time of vehicles in the Far-field or MIT scenes usually lasts longer than 10 seconds, vehicle activities are now captured as a single motif as shown in Fig. 18 rather than as a sequence of shorter motifs of 5 to 10 seconds in length. As an example, the motif in Fig. 18(a) combines the activities of Fig. 15(e,d). The same applies with the pedestrian activities in the Traffic Junction case (cf Fig. 18(i,j)).

[7] Note however that longer motifs increase the chance of observing some random co-occurrences, as the amount of overlap with other activities, potentially unexplained by current motifs, increases as well. This is particularly true when the amount of data is not very large like in the Traffic junction case.



(a) $p(z) = 0.090$ (b) $p(z) = 0.089$ (c) $p(z) = 0.085$ (d) $p(z) = 0.073$

(e) $p(z) = 0.058$ (f) $p(z) = 0.052$ (g) $p(z) = 0.052$ (h) $p(z) = 0.048$

Fig. 17: Traffic Junction data. Representative motifs of 10s duration. (a-d) vehicle activities. (e-h) pedestrian activities. Best viewed in color. Please see the page at [1] to view animated GIF versions of the motifs.

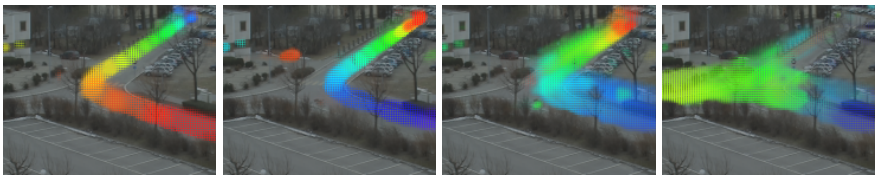
6.3 Event detection

To evaluate how well the recovered motifs match the real activities observed in the data, we performed a quantitative analysis by using the PLSM model to detect particular events. Indeed, as the model can estimate the most probable occurrences $p(t_s, z|d)$ of a topic z for a test document d, it is possible to 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 between precision and completeness (i.e. recall).

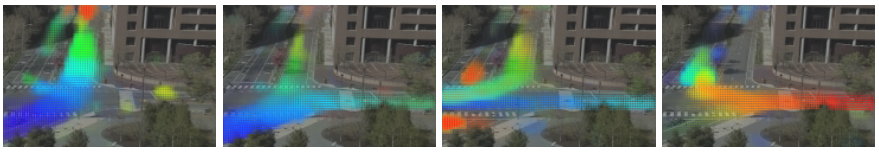
For this event detection task, we labeled a 12 minute video clip from the Far-field scene, distinct from the training set, and considered all the different car activities that pass through the three road junction. Activity categories that occurred fewer than 5 times in this test data were discarded, which left us with the 3 activity categories depicted in Fig. 19 with a total of 51 occurrences. For each ground truth category, we manually associated one of the discovered motifs of maximum 10 second duration[8]. The motifs considered for event detection are shown in 15(a,d,i). Using the occurrences $p(t_s, z|d)$ of these motifs, precision/recall curves were computed. They are shown in Fig. 19.

From the curves, it is evident that for two out of the three events, we obtain a close to 100% result. Indeed, the worst performance is for the activity "top right to bottom right" which gives a precision above 80% for a very high recall of 90%. This proves that the discovered motifs match the real scene activities well, and that motif starting times could be exploited for real event detection.

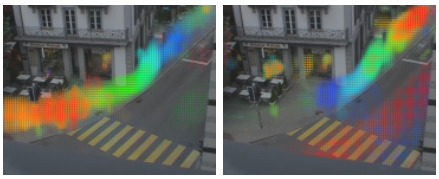
[8] To perform the association we allowed a constant offset between the event in the ground truth, and the starting time of a motif learned from PLSM.



(a) $p(z) = 0.101$ (b) $p(z) = 0.096$ (c) $p(z) = 0.036$ (d) $p(z) = 0.049$



(e) $p(z) = 0.092$ (f) $p(z) = 0.066$ (g) $p(z) = 0.060$ (h) $p(z) = 0.055$



(i) $p(z) = 0.102$ (j) $p(z) = 0.088$

Fig. 18: Motifs of 20s duration that mainly differ from their 10s shorter counterparts. (a–d) Far-field, (e–h) MIT, and (i–j) Traffic junction data. All the above motifs capture the full extent of the activities within the scene. Best viewed in color. Please see the page at [1] to view animated GIF versions of the motifs.

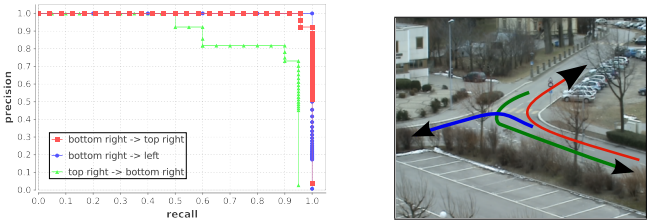


Fig. 19: Precision/recall curves for the detection of 3 types of events mapped onto 3 topics, evaluated on a 12 minute test video.

6.4 Activity prediction

The predictive model. The learned PLSM model can be used for predicting the most probable future words. We have thus defined our task as estimating the probability $p_t^{pred}(w)$ that a word w appear at time t given all past information, that is, given the temporal document $n(w, t_a, d)$ up to time $t_a = t - 1$.

In our generative modeling approach, a word at time t can occur due to either a motif that has already started at a past time $t_s \in [t - T_z + 1, t - 1]$, or due to a

motif that starts at the same time t. Hence, we define the prediction model as:

$$p_t^{pred}(w) \propto (1-\gamma) \sum_{t_s=t-T_z+1}^{t-1} \sum_z \hat{p}(t_s, z|d) p(w, t-t_s|z) + \gamma \sum_z p(z) p(w, 0|z), \quad (17)$$

where $\hat{p}(t_s, z|d)$ denotes our estimation that the topic z starts at time t_s given the observed data, γ represents the probability that a topic starts at the current instant, and $p(z)$ represents the motif prior probability estimated (along with the motifs) on training data[9]. To set γ, we have given equal priority to the starting time instants, and set $\gamma = \frac{1}{T_z}$, i.e. a value of 0.1 in the current experiments. To obtain $\hat{p}(t_s, z|d)$ we simply apply our inference procedure to the temporal document $n(w, t_a, d)$ using only observations up to time $t-1$ and re-normalize the resulting $p(t_s, z|d)$ so that $\sum_{t_s=t-T_z+1}^{t-1} \sum_z \hat{p}(t_s, z|d) = 1$.

Evaluation protocol and results. The model was evaluated as follows on the MIT and Far-field datasets. The motifs and motif prior $p(z)$ were learned using 90% of the data and tested on the remaining 10%. This resulted in 4900 and 5900 time steps (seconds) for training and $N_{test} = 550$ (9 mins) and $N_{test} = 720$ (12 mins) time steps for testing in the MIT and Far-field cases respectively. The performance of the task was measured by using the average normalized prediction log-likelihood defined as:

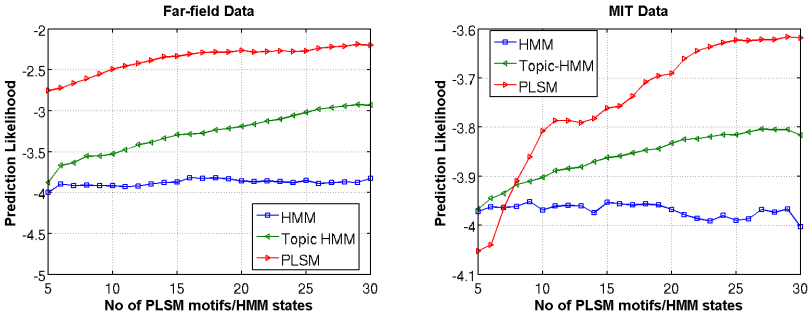
$$ANL = \frac{1}{N_{test}} \sum_t \frac{\sum_w n(w, t, d) \log(p_t^{pred}(w))}{\sum_w n(w, t, d)} \quad (18)$$

The ANL measure is a standard performance evaluation measure used in evaluating the model performance [35,31]. It is also inversely related to the perplexity measure that is used in topic models [12,4]. A higher value for ANL indicates a better predictive capacity and vice versa. In order to compare the prediction accuracy of our model, we implemented two other temporal models.

Simple HMM. Here, the sequences of observation vectors $o_t(w) = n(w, t, d)$ from the training temporal documents were used to learn in an unsupervised fashion (i.e. by maximizing the data-likelihood) a fully-connected HMM with n states. The emission probabilities were defined as Gaussians with a diagonal covariance matrix. At test time, the trained HMM was used to compute the expected state probability at time t given all observations up to time $t-1$, from which the expected observation vector (and hence a predicted word probability $p_t^{pred}(w)$) was inferred.

Topic HMM. The second model is a more sophisticated approach in line with [13], wherein the Markov chain models the dynamics of a global behavior state. More precisely, we first apply PLSA (with n topics) to the set of training documents $\{o_t, t \in training\}$. This results in a set of topics $p(w|z)$ and topic distributions $o'_t(z) = p(z|o_t)$. We then learn an HMM with n states using the topic observation sequence o'_t. The HMM states learned with this method capture distinct scene level behaviors characterized by interacting topics and the Markov chain models the temporal dependencies among them. We thus refer to this method as *Topic*

[9] Note that rather than simply using $p(z)$ as the prior for a topic to start at time t, we could have further exploited the past informations available in the past motif occurrences $\hat{p}(t_s, z|d)$ (e.g. the motif of Fig. 15(e) is often followed by that of Fig. 15(d) several seconds later). However, as this is not part of our model, we preferred to go for the simpler case.



(a) Far-field data (b) MIT data

Fig. 20: Average Normalized Prediction log-likelihoods for (a) Far-field data, (b) MIT data. In both plots, the x-axis represents either the number of motifs (PLSM model), or the number HMM states in the two other cases.

HMM. At test time, the expected state, topic and word probability distributions can be successively computed using the learned model.

Fig. 20 presents the results of PLSM and the two competitive methods. We observe that the simple HMM method gives the worst predictions on both datasets compared to the more sophisticated Topic-HMM, whose observations come from the PLSA topics. However, overall, the PLSM model gives a much better performance than the two HMM based methods, showing that the incorporation of temporal information at the topic level rather than at the global scene level is a better strategy.

In the Far-field case, where the scene is not governed by any specific rules, PLSM performs consistently and significantly better with an average likelihood of almost one order of magnitude greater than the Topic-HMM (ANL of around -2.15 compared to -2.9 for the Topic-HMM when $n = 30$). On the MIT data, the situation is somewhat different. When the number of states/motifs is low (until $n = 8$), the HMM approaches are performing better as they are more able to model the different phases of the regular cycle governed by the traffic lights that the scene goes through. These distinct global behavior states, and the transitions between them, are captured explicitly in the Topic-HMM and to a lesser extent in the HMM method whereas our method does not have any prior on the sequences of motif occurrences. Nevertheless, as the number of motifs increases, PLSM provides a finer and more detailed description of the activities and its prediction accuracy improves beyond the performance of the other methods that have difficulties to take advantage of the modeling of questionable and unpredictable sub-phase global scene activity patterns. Note however that the difference with the other model is not as high in this case as on the Far-field data. Finally, it is interesting to note that the prediction accuracy of the PLSM method tends to saturate for a number of motif N_z close to that selected using the BIC criterion (20 for the Far-field data, 26 for MIT data).

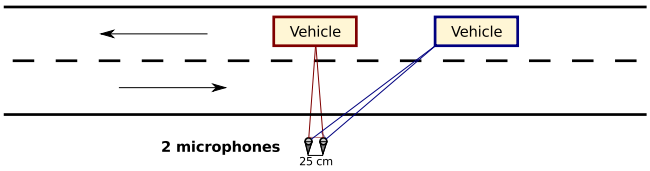


Fig. 21: Audio scene analysis setup. An array of two microphones is located on a road side. The audio time difference of arrival (TDOA) between these microphone is measured and provides information about the azimuths of sound sources.

7 Audio Scene Analysis with Microphone array

The PLSM model can be applied to any multivariate time-series that can be described as word×time document. To test the generality of the model, we used it for analysing a scene using acoustic data. In the following we describe the set-up and data that we used, and then present our results.

The setup is described in Fig. 21. The recording was done using two microphones located on the side of a two way road where the main activities are essentially vehicles either going from left to right or from right to left, at different speeds. In this experiment, our activity feature characterizes the sound source locations, and relies on the time difference of arrival (TDOA) principle: a sound generated by a source located at an azimuth angle θ relative to the microphone pair arrives at the microphones with a time difference of $\tau(\theta)$ between them. Thus, to build the temporal document for PLSM, we use dense TDOA information extracted from the microphones as follows. At each time instant t_a, we compute on an 80ms temporal window the generalized cross-correlation $GCC(\tau)$ between the two signals for different τ values corresponding to azimuth angles from almost -90° to 90°. We then normalize the measurements, and further subtract a uniform value from the result. The normalization provides some invariance to car loudness, while the subtraction removes uniform noise that might have been amplified by the normalization step. Finally, the representation is simplified by averaging the measurements on 25 regular intervals $\Delta\tau_i$ to measure the "amount" of sound signal coming from the direction $\theta(\tau_i)$ and construct the word-time frequency matrix $n(w_{\tau_i}, t_a, d)$. Fig. 22 shows a sample document (a clean one) with multiple vehicles passing: five cars going from left to right (upward ramp) and one car going from right to left (downward ramp).

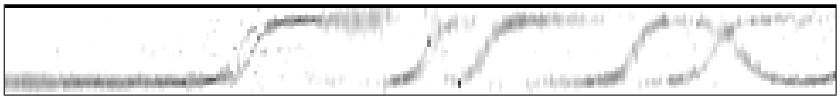


Fig. 22: TDOA sample temporal document showing multiple occurrences of cars going from left to right (upward ramp) and one occurence of car going from right to left (downward ramp) overlapping. The horizontal axis is time (one time step is 80ms), the vertical axis is the azimuth angle.

Fig. 23: Four sequential motifs of 30 timesteps ($\approx$ 2.5 seconds) from TDOA data.

Fig. 24: Four sequential motifs of 60 timesteps ($\approx$ 5 seconds) from TDOA data.

For the experiments, 30 recordings of approximately 20 seconds each were used, comprising a total of around 120 car passing events. Given the 80ms time step, each recording produced a temporal document of around 250 time instants with 25 possible words (angles). The PLSM approach was applied to these documents, with the same MAP setting and sparsity level ($\lambda = 0.25$) as in the video case. However, given the known expected number of topics (four), we did not use the BIC criterion. The results are shown in Fig. 23 when using a maximum length of 30 time steps ($\approx$ 2.5 seconds). Despite the noise and variations in vehicle speed (from around 35 to 70km/h), we observe that the dominant patterns are clearly captured: the ramp ones, corresponding to the car passing in front of the microphones; and the almost stationary motifs corresponding to cars approaching or leaving. Indeed, in this latter cases, azimuth angles are around $+90$ or -90 degrees and do not vary much. These activities get captured as separated motifs (from the ramp ones) because the measured duration of the "approach phase" is highly variable and depends on the sound volume of the car: a louder car will be perceived earlier by the microphones. Similar results were obtained when searching for motifs from 30 to 70 time steps. For instance, Fig. 24 shows the results with a length of 60 time steps($\approx$ 5 seconds).

8 Conclusion

In this paper we proposed a novel unsupervised approach for discovering dominant activity motifs from multivariate temporal sequences. Our model infers temporal patterns of a maximum time duration by modeling the temporal co-occurrence of visual words, which significantly differs from previous topic model based approaches. This is made possible thanks to the introduction of latent variables representing the motif start times, bringing the following advantages: a) they help in implicitly aligning occurrences of the same motif while learning, and b) they allow us to infer when an activity starts. The model parameters can be inferred efficiently using en Expectation-maximization procedure that exploits a novel sparsity constraint. The effectiveness of our model was extensively validated using synthetic as well as multi-modal real life data sets from both the visual and acoustic domains. Qualitative results and quantitative experiments on event detection and prediction tasks showed that the approach was discovering motifs consistent with the scene activities and was resulting in superior performance compared to other state of the art Dynamic Bayesian Network based alternatives.

The model offers room for further improvements. For instance, although we have used the Bayesian Information Criteria measure to determine the number of topics, we still observe a few motifs (usually of lower $p(z)$) that are copies or minor variations of other motifs, which could hence be merged. We believe that this could be better dealt with by using other data driven approaches like [28], or by explicitly disfavoring the recovery of similar motifs. Similarly, while our model handles local variations in local activity execution timing well, it can only cope to a certain extent with differences in overall execution speed. There are several ways to handle this. First, we can conduct an a-posteriori analysis, by identifying motif replicas differing by speed execution variations. Or one can introduce an explicit latent variable to model the execution speed. Although this can be added in a straightforward manner in our model, this would result in increased computational complexity. Finally, our model identifies activities and their starting times, but has no higher-level representation of the motif occurrences. The analysis and modeling of these occurrences in terms of dependencies or interactions could enhance the global understanding of the scene through, for instance, the identification of scene level rules (e.g. right of way) or activity cycles due to the presence of a traffic light.

Acknowledgements The authors thank Patrick Marmaroli, EPFL for providing the TDOA data and Carl Scheffler for his useful comments.

References

1. `http://www.idiap.ch/~vjagann/plsm.html`
2. Blei, D., Lafferty, J.: A correlated topic model of science. Annals of Applied Statistics **1**(1), 17–35 (2006)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: International Conference on Machine Learning, pp. 113–120 (2006)
4. Blei, D.M., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research (3), 993–1022 (2003)
5. Boiman, O., Irani, M.: Detecting irregularities in images and in video. International Journal of Computer Vision **74**(1), 17–31 (2007)
6. Chien, J.T., Wu, M.S.: Adaptive bayesian latent semantic analysis. IEEE Transactions on Audio, Speech, and Language Processing **16**(1), 198–207 (2008)
7. Faruquie, T.A., Kalra, P.K., Banerjee, S.: Time based activity inference using latent Dirichlet allocation. In: British Machine Vision Conference. London, UK (2009)
8. Girolami, M., Kabán, A.: On an equivalence between PLSI and LDA. In: ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 433–434 (2003)
9. Gohr, A., Hinneburg, A., Schult, R., Spiliopoulou, M.: Topic evolution in a stream of documents. In: SIAM International Conference on Data Mining, pp. 859–870 (2009)
10. Gruber, A., Rosen-Zvi, M., Weiss, Y.: Hidden topic Markov model. In: International Conference on Artificial Intelligence and Statistics. San Juan, Puerto Rico (2007)
11. Hervieu, A., Bouthemy, P., Cadre, J.P.L.: A statistical video content recognition method using invariant features on object trajectories. IEEE Transactions on Circuits and Systems for Video Technology **18**(11), 1533–1543 (2008)
12. Hofmann, T.: Unsupervised learning by probability latent semantic analysis. Machine Learning **42**, 177–196 (2001)
13. Hospedales, T., Gong, S., Xiang, T.: A Markov clustering topic model for mining behavior in video. In: IEEE International Conference on Computer Vision. Kyoto, Japan (2009)
14. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research **5**(2), 1457–1470 (2005)
15. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Journal of Knowledge and Information Systems pp. 263–286 (2000)

16. Kuettel, D., Breitenstein, M.D., Gool, L.V., Ferrari, V.: What's going on? discovering spatio-temporal dependencies in dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1951–1958 (2010)
17. Li, J., Gong, S., Xiang, T.: Global behaviour inference using probabilistic latent semantic analysis. In: British Machine Vision Conference (2008)
18. Li, J., Gong, S., Xiang, T.: 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)
19. Luvison, B., Chateau, T., Sayed, P., Pham, Q.C., Laprest, J.T.: An unsupervised learning based approach for unexpected event detection. In: International Conference on Computer Vision Theory and Applications (VISAPP), Lisboa, pp. 506–513 (2009)
20. Makris, D., Ellis, T.: Automatic learning of an activity-based semantic scene model. IEEE International Conference on Advanced Video and Signal Based Surveillance **2**(1), 183–188 (2003)
21. Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B.: Exact discovery of time series motifs. In: SIAM International Conference on Data Mining, pp. 473–484 (2009)
22. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3), 299–318 (2008)
23. Quelhas, P., Monay, F., marc Odobez, J., Gatica-perez, D., Tuytelaars, T.: A thousand words in a scene. IEEE Transactions on Pattern Analysis and Machine Intelligence (2005)
24. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)
25. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: IEEE International Conference on Computer Vision (2005)
26. Stauffer, C., L.Grimson, E.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 747–757 (2000)
27. Tanaka, Y., Iwamoto, K., Uehara, K.: Discovery of time-series motif from multi-dimensional data based on MDL principle. Machine Learning **58**, 269–300 (2005)
28. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association **101**(476), 1566–1581 (2006)
29. Tommasi, C., Kanade, T.: Detection and tracking of point features. International Journal of Computer Vision (1991)
30. Varadarajan, J., Emonet, R., Odobez, J.: Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. In: British Machine Vision Conference, pp. 117.1–117.11. Aberystwyth (2010)
31. Varadarajan, J., Odobez, J.: Topic models for scene analysis and abnormality detection. In: IEEE International Workshop on Visual Surveillance. Kyoto, Japan (2009)
32. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: International Conference on Machine Learning, pp. 977–984. Pittsburgh, Pennsylvania (2006)
33. Wang, C., Blei, D.: Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In: Neural Information Processing Systems, pp. 1982–1989 (2009)
34. Wang, C., Blei, D.M., Heckerman, D.: Continuous time dynamic topic models. In: Conference on Uncertainty in Artificial Intelligence (2008)
35. Wang, X., Ma, X., Grimson, E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(3), 539–555 (2009)
36. Wang, X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: ACM Conference Knowledge Discovery and Data Mining. Philadelphia, USA (2006)
37. Wang, X., Tieu, K., Grimson, E.L.: Learning semantic scene models by trajectory analysis. In: European Conference on Computer Vision, vol. 14, pp. 234–778 (2004)
38. Williamson, S., Wang, C., Heller, K., Blei, D.: Focused topic models. In: NIPS workshop on Applications for Topic Models: Text and Beyond. Whistler, Canada. (2009)
39. Xiang, T., Gong, S.: Video behavior profiling for anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(5), 893–908 (2008)
40. Yang, Y., Liu, J., Shah, M.: Video scene understanding using multi-scale analysis. In: IEEE International Conference on Computer Vision. Kyoto, Japan (2009)
41. Yi Zhang Jeff Schneider, A.D.: Learning compressible models. In: Proceedings of SIAM Data Mining (SDM) Conference (2010)

42. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., G.Lathoud: Multimodal group action clustering in meetings. In: ACM International Conference on Multimedia, Workshop on Video Surveillance and Sensor Networks (2004)
43. Zhong, H., Jianbo, S., Mirko, V.: Detecting unusual activity in video. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 819–826. Washington, DC (2004)