Temporal Analysis of Motif Mixtures using Dirichlet Processes

Rémi Emonet, *Member, IEEE,* Jagannadan Varadarajan, *Member, IEEE*
and Jean-March Odobez, *Member, IEEE*

Abstract—In this article, we present a new model for unsupervised discovery of recurrent temporal patterns (or motifs) in time series (or documents). The model is designed to handle the difficult case of multivariate time series obtained from a mixture of activities, that is, our observations are caused by the superposition of multiple phenomena occurring concurrently and with no synchronization. The model uses non parametric Bayesian methods to describe both the motifs and their occurrences in documents. We derive an inference scheme to automatically and simultaneously recover the recurrent motifs (both their characteristics and number) and their occurrence instants in each document. The model is widely applicable and is illustrated on datasets coming from multiple modalities, mainly, videos from static cameras and audio localization data. The rich semantic interpretation that the model offers can be leveraged in tasks such as event counting or for scene analysis. The approach is also used as a mean of doing soft camera calibration in a camera network. A thorough study of the model parameters is provided and a cross-platform implementation of the inference algorithm will be made publicly available.

Index Terms—motif mining, mixed activity, unsupervised activity analysis, topic models, multi-camera, camera network, non parametric models, Bayesian modeling, multivariate time series

1 Introduction

Mining recurrent temporal patterns in time series is an active research area. The objective is to find, with as little supervision as possible, the recurrent temporal patterns (or motifs) in multivariate time series. One major challenge comes from the fact that time series that we observe are often "caused" by a superposition of different phenomena recurring in time and with no specific synchronization.

This problem has its instances in many domains and we can illustrate this with one of the application we use in this article: activity extraction from video. In a video sequence, what we observe is the result of different activities acted by different persons or objects present in the scene. In such cases, we would use long term recordings to learn the independent activities and spot their occurrences automatically.

Many other time series can present the same characteristic of being a fusion of multiple activities. For example, we can consider the time series made of the overall electric and water consumption of a building. In such setting, we could observe activity patterns (motifs) like a short water consumption followed by short electric consumption (someone filling and then starting a boiler). We could also observe motifs like alternating water and electric consumptions for one hour (for a washing machine cycle). As multiple persons can live in the building (and one person can also do multiple tasks), multiple occurrences of these two motifs can occur at the same time and with no specific synchronization.

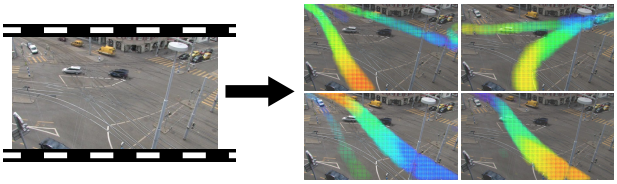


Fig. 1. Task on video sequences. Without supervision, we want to extract recurring temporal activity patterns (4 are shown here). Time is represented using a gradient of color from blue to red. We call these patterns "*motifs*" in the article, the representation is explained in Fig. 10.

In the context of video sequences, the specific goal is to find activity patterns (e.g. car passing, pedestrian crossing) without supervision. This elementary task can be useful for applications like summarizing a scene, counting or detecting particular events or detecting unusual activities. More generally, this identification of temporal motifs and the instant at which they occur can be used as a dimensionality reduction method for (potentially supervised) higher level analysis.

In this paper, we present a model for finding temporal patterns (motifs) in time series. While selecting the number of motifs automatically, we also determine the number of times they occur in the data, and when they occur. We also propose an extension of the model that automatically infers the length of the motifs.

Compared to our previous work in [1], this paper presents many additional elements including both increased details and newer contributions:

- *Contact:* remi@heeere.com vjagan@gmail.com odobez@idiap.ch
- *Authors were all working at the IDIAP Research Institute.*

- a new model which allows for variable size motifs;
- a clarified methodological section;
- a more detailed study of the effect of parameters;
- a fully revisited experimental section;
- a calibration-free, multicamera experiment.

The proposed models are well suited for any time series generated by non-synchronized concurrent activities as we illustrate by applying it to real video sequences.

In Section 2, we introduce approaches that are related to the one proposed in this article which is quickly introduced in Section 3. The full details of the proposed models are then provided in Section 4. A discussion about the meaning of the parameters of the models is given in Section 6, followed by details about the inference procedure in Section 5. In sections 7 and 8, we present our experiments on different kinds of data including synthetic data, traffic video data, multi-camera video surveillance data and audio data. Section 7 and 8 also provides a comparison with other methods from the state of the art. We finally discuss our contribution and conclude in Section 9.

2 Related Work

Unsupervised activity modeling – Recently, there has been an increased focus on discovering activity patterns from videos, especially in surveillance scenarios. These patterns are often called "activities" (or "motifs") in the existing literature. Although other paradigms can be successful as well (e.g. see an approach based on diffusion maps for instance [2]), topic models have shown tremendous potential in achieving this in an unsupervised fashion. Most of the existing topic model based methods propose to break the videos into clips of a few frames or seconds. Documents are created from these clips by quantizing pixel motion at different locations in the images. This approach was followed in [3], [4], [5], where activities are represented as static co-occurrences of words.

Activities in a video are by nature, temporally ordered. Therefore, representing each action as a bag-of-words as in [3], [4] results in loosing temporal dependencies among the words. Several attempts have been made to incorporate temporal information in topic models, starting from the work done in text processing [6], [7]. Following these lines, the method from [8] improves by modeling the sequence of scene behaviors as a Markov model, but with a pre-determined fixed set of topics. While the temporal order is imposed at the global scene level, the higher level of the hierarchy, the activity patterns are still modeled as static distributions over words.

The methods proposed in [9], [10], [11] complement visual words with their time stamps to recover temporal patterns. While this method can be useful when clips are aligned to recurring cycles like traffic signals (as this was done manually in [10]), it gives poor results in general cases where such alignment is not done a priori as in [9]. Illustrations and more insights about how these methods operate can be found in the experimental section (Section 7.3) of this paper. A more general approach was proposed in [11], wherein motifs and their starting times are jointly learnt, requiring no manual alignment of clips. However, the model is not fully generative and requires setting various parameters like the number of topics.

One of the main challenges in topic model based activity modeling is model selection, that is, the automatic estimation of the number of topics. Non-parametric Bayesian methods such as Hierarchical Dirichlet Process [12] allows, in theory, to have an infinite number of topics, and in practice, to select this number automatically. Such a model was explored for discovering static topics in works such as [3], [13] and [14].

In order to integrate temporal information in HDP, both [5] and [15] use the infinite state HDP-HMM paradigm of [12] to identify temporal topics and scene level rules. Even if the approach proposed in [5] is designed to find multiple HMMs to model different local scene rules, in practice, only a single HMM was found for all tested scenes. The single HMM that is recovered eventually captures the rules at a global scene level, similarly to what was done in [8]. In [5], the failure to capture local activities as expected might be due to the loosely constrained model that they propose: the considered HMMs are fully connected and the model does not try to explicitly find the starts of activity. This results in a model with too many freedom and makes it highly improbable that inference will converge to a result involving multiple local rules.

Multi-camera analysis – The flexibility of the approach presented in this paper make it possible to apply it to a multi-camera setup. Whereas many works require to provide both intrinsic camera parameters and inter-camera calibrations, works more related to our approach investigate the recovery of the topology of a set of cameras as in [16]. Most of these methods rely on per-camera tracking combined with re-identification from camera to camera in order to infer the layout of the camera network. Person re-identification is usually carried out using joint appearance matching (e.g. color) and inter-camera travel time modeling. However, as motivated for example in [17], videos from crowded public spaces like those installed in metro stations have very poor quality, low frame rate and suffer from a lot of occlusions. In such setup, robust person tracking is still an unsolved issue and thus extracting distinctive features for re-identification is ineffective.

As an alternative, authors in [17] rely on more robust, lower level activity feature (background subtraction) computed over data-driven region segments extracted independently for each image. Cross Canonical Correlation Analysis (xCCA) is then applied to extract relations between all pairs of regions in all views and derive the topology of the camera network. Although the obtained model is used to improve person re-identification across views, the approach does not result in a detailed tempo-

ral activity model with automatic soft calibration (even when views overlap) as we propose.

Simple topic models like Latent Dirichlet Allocation (LDA) have also been used for co-occurrence analysis in order to capture the topology of a camera network. In [18] and [19], an ad hoc model that improves over LDA by adding some side information is used to find camera links. By encouraging observed trajectories that are close enough in time to have comparable topic distributions, the method captures latent topics that correspond to multi-camera activities. As it relies on rather clean urban trajectories and does not try to explicitly unmix the activities (all pairs of close trajectories are linked, whether they correspond or not) this method exhibits some limitations for crowded scenes.

Major contributions – Our paper differs significantly from the approaches presented above. Our aim is to find both motifs with strong explicit temporal information and when they appear in the temporal documents. The major contributions of this article include:

- the application of non-parametric Bayesian principles to temporal topic modeling to automatically determine the number of motifs shared by the documents and also find when they appear in each temporal document;
- the introduction of a model that is able to find the length of each motif;
- the derivation of a Gibbs sampler for the joint inference of the topics and their occurrences;
- extensive experiments on a large set of video data provided by [3], [5], [8], [11], [20] and comparison with other approaches;
- the application of our model to a mutli-camera setup and to binaural audio data.

3 Approach Overview

The input to our method is a set of temporal documents (possibly a long single one) as presented in section 1. This observed document is defined as a table of counts, where the entries reflect the amount of presence of a word from a fixed vocabulary at every instant of the temporal document. Our approach is depicted in figure 2 where each document is represented as a set of "motif occurrences" (e.g., 7 of them in Fig. 2). Each occurrence is defined by a starting time instant and a motif. Motifs are shared by different occurrences within and across documents.

In our model, an important aspect is the use of Dirichlet Processes (DP). A DP is a non-parametric Bayesian process that represents an infinite mixture model (see Section 4.1 for details). The term "non-parametric" refers to the fact that the model grows automatically to account for the observations. Dirichlet processes are often used to determine automatically the number of relevant elements in a mixture model (e.g., number of topics or number of gaussians). A DP is an infinite mixture but

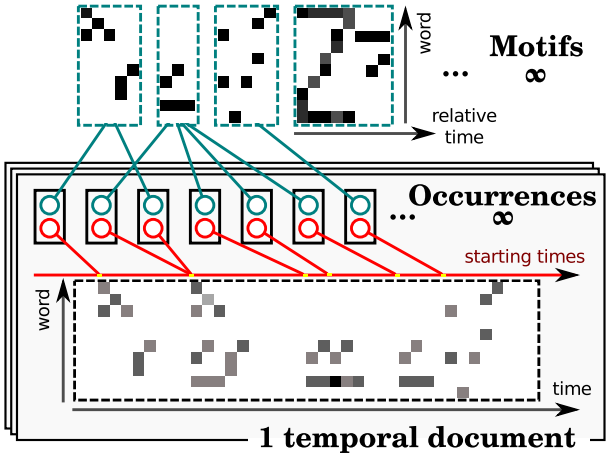


Fig. 2. Schematic generative model. A temporal document is made of word counts at each time instant. Each document is composed of a set of occurrences, each being defined as a motif type and a starting time. The motifs are shared by the occurrences within and across documents.

observations from a DP most probably tend to cluster on some limited elements of the mixture.

We use two levels of DP in our approach. At a lower level, within each document, we model the set of occurrences using a DP: the observations then cluster around an automatically determined number of occurrences. At a higher level, we model the set of motifs using a DP: the occurrences (and their associated observations) within and across documents then cluster around an automatically determined number of motifs. With this hierarchical approach, each observation is associated through its occurrence to a motif.

In this article, we introduce a model that generalizes the one proposed in [1] where motifs have a fixed length. In this improved model, the settings of the hyper-parameters enable us to manipulate the length of the motifs. In particular, it allows us to either use a fixed motif length or find the lengths of the motifs automatically.

4 Proposed Model

As introduced in section 3, our model relies on Dirichlet Processes (DP) to discover motifs, their number, and find their occurrences. We will thus start by introducing DP before describing the core of our model in details. Then, we will explain the two variations of our model called Temporal Analysis of Motif Mixtures (TAMM) and Variable Length TAMM (VLTAMM).

4.1 Background on Dirichlet Processes (DP)

Here we introduce Dirichlet Processes, a method to naturally handle infinite mixture models and a building block of our proposed models. The mixture components we are

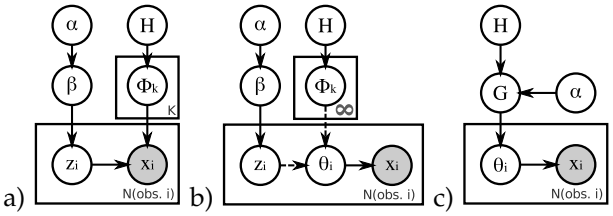


Fig. 3. Finite mixture and Dirichlet Process (infinite mixture): a) finite mixture with K elements; b) mixture representation for DP; c) compact representation for DP.

using are categorical distributions[1] but all elements in the current subsection can be interpreted identically with any mixture model such as a Gaussian Mixture Model. We use $Comp$ to denote the component distribution in this introductory section.

Fig. 3a is a graphical representation of a finite mixture model with K components. The β vector is giving the weight of each mixture component and α is a prior (possibly uninformative) on these weights. Each Φ_k represents the parameters of a mixture component and for each observation x_i, z_i represents the index of the mixture component this observation is coming from.

Fig. 3b first shows that we can explicitly represent the mixture component selected by each observation noted θ_i. We use dashed arrows to indicate deterministic relations, here $\theta_i = \Phi_{z_i}$ (or, expressed as a draw from a Dirac distribution: $\theta_i \sim \delta_{\Phi_{z_i}}$). More importantly, Fig. 3b also illustrates the uniqueness of a Dirichlet Process, i.e., there are an infinite number of mixture components instead of a finite number K. To adapt to this infinite mixture elements, the weight vector β is of infinite length and the prior α takes a specific form. The α prior is now a single positive real value used as the parameter of a "GEM" (Griffiths, Engen, McCloskey) also known as a "stick breaking" process. This process produces an infinite list of weights that sum to 1: the first weight $\beta_1 = \beta'_1$ is drawn from a beta distribution $Beta(1, \alpha)$, the second weight is drawn in the same way but only from the remaining part, i.e. $\beta_2 = (1 - \beta_1) * \beta'_2$ with β'_2 drawn from $Beta(1, \alpha)$, and so on for the other weights, hence the "stick breaking" name. In addition to these weights, each mixture component parameter set Φ_k is drawn independently from a prior H, and each observation is drawn from its mixture component. We thus have:

$$\beta \sim GEM(\alpha) \quad (1)$$

$$\forall k \qquad \phi_k \sim H \quad (2)$$

$$\forall i \qquad z_i \sim \text{Categorical}(\beta) \quad (3)$$

$$x_i \sim Comp(\phi_{z_i}) \quad (4)$$

A more compact equivalent notation can be used to represent a Dirichlet Process. While the mixture representation is well adapted for deriving the Gibbs sampling scheme, the compact representation is widely used and might help us to get a quick overview of the model. In the compact representation from Fig. 3c, individual mixture components are not shown and instead their weighted countable-infinite mixture $G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ is used. The corresponding representation, using a DP notation, is given as:

$$G \sim DP(\alpha, H) \quad (5)$$

$$\forall i \qquad \theta_i \sim G \quad (6)$$

$$x_i \sim \theta_i \quad (7)$$

1. sometimes "multinomial" is used in place of "categorical"

4.2 Base of the Proposed Model

Our goal is to automatically infer a set of motifs (temporal activity patterns) from a set of documents containing time-indexed words.

More precisely, let us define a document j as a set of observations $\{(w_{ji}, at_{ji})\}_{i=1...N_j}$, where w_{ji} is a word belonging to a vocabulary $\mathcal{V}$ and at_{ji} is the absolute time instant at which the observation occurs within the document. For instance, in case of a video, each word of the vocabulary is describing a spatially localized motion in the image (how we get these words is defined in the experiment section).

We also consider time information when defining our "motifs" as temporal probabilistic maps. More precisely, if ϕ_k denotes a motif table (i.e., the parameters of a categorical distribution), then $\phi_k(w, rt)$ denotes the probability that the word w occurs at a relative time instant rt after the start of the motif.

Our goal is to infer the set of motifs from one or more temporal documents. As discussed previously, this must be done altogether with inferring the occurrences (instants of occurrence) of all motifs in the documents. As it is difficult to fix the number of motifs before hand, we use a DP to allow the learning of a variable number of motifs from the data. Similarly, within each temporal document, we use another DP to model all motif occurrences as we don't know their number in advance.

Our generative model is thus defined using the graphical models presented in Figure 4. Fig. 4a depicts our model using the compact Dirichlet process notation as done for DP in Fig. 3c, whereas Fig. 4b depicts the developed notation (cf Fig. 3b). Notice that in these drawings, two variables in an elongated circle form a couple, indicating that they are generated together: the pair itself is drawn from a distribution over the pairs.

The following describes more the model which involves a number of variables. The interested reader can refer to Section 6 for a second explanation of the meaning of the model's hyper-parameters.

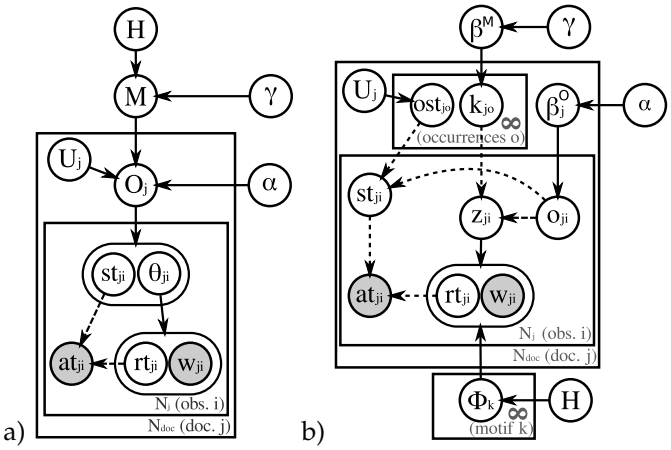


Fig. 4. Proposed model a) with DP compact notation; b) with developed Dirichlet processes (using stick-breaking convention at both levels). Dashed arrows represents deterministic relations (conditional distributions are a Dirac).

The equations associated with Fig. 4a are as follows:

$$M \sim DP(\gamma, H) \text{ where } H = Dir(\eta) \quad (8)$$
$$\forall j \qquad O_j \sim DP(\alpha, (U_j, M)) \quad (9)$$
$$\forall j \forall i \qquad (st_{ji}, \theta_{ji}) \sim O_j \quad (10)$$
$$(rt_{ji}, w_{ji}) \sim \text{Categorical}(\theta_{ji}) \quad (11)$$
$$at_{ji} = st_{ji} + rt_{ji} \quad (12)$$

where deterministic relations are denoted with "=". The first DP level generates our list of motifs in the form of an infinite mixture M. Each of the motif is drawn from H, defined as a Dirichlet distribution of parameter η (a table of the size of a motif; see section 4.3 for how we set it).

Contrary to simpler mixture models such as LDA or HDP, our set of mixture components is not only shared across documents, but also across motif occurrences using the DP at the second level. More precisely, the document specific distribution O_j is not defined as a mixture over motifs, but as an infinite mixture over occurrences from "start-time $\times$ motif" (cf Fig. 2), since the base distribution is defined by (U_j, M). Each of the atoms is thus a couple (ost_k, ϕ_k), where $ost_k \sim U_k$ is the occurrence starting time drawn from U_j, a uniform distribution over the set of possible motif starting times in the document j, and $\phi_k \sim M$ is one of the topic drawn from the mixture of motifs.

Observations (w_{ji}, at_{ji}) are then generated by repeatedly sampling a motif occurrence (Eq. 10), using the obtained motif θ_{ji} to sample the word w_{ji} and its relative time in the motif rt_{ji} (Eq. 11). From the relative time rt_{ji}, using the sampled starting time st_{ji}, the word absolute time occurrence at_{ji} can be deduced (Eq. 12).

The fully developed model given in Fig. 4b helps to understand the generation process and the inference better. The corresponding equivalent equations can be written as:

$$\beta^M \sim GEM(\gamma) \quad (13)$$
$$\forall k \qquad \phi_k \sim H \quad (14)$$
$$\forall j \qquad \beta^o_j \sim GEM(\alpha) \quad (15)$$
$$\forall j \forall o \qquad ost_{jo} \sim U_j \quad \text{and} \quad k_{jo} \sim \beta^M \quad (16)$$
$$\forall j \forall i \qquad o_{ji} \sim \beta^o_j \quad (17)$$
$$z_{ji} = k_{jo_{ji}} \quad \text{and} \quad st_{ji} = ost_{jo_{ji}} \quad (18)$$
$$(rt_{ji}, w_{ji}) \sim \text{Categorical}(\phi_{z_{ji}}) \quad (19)$$
$$at_{ji} = st_{ji} + rt_{ji} \quad (20)$$

The main difference with the compact model is that the way motif occurrences are generated is explicitly represented. Occurrences are the analog of the "tables" in the Chinese Restaurant Process analogy of the HDP model: both the global GEM distribution over motifs β^M and U_j are used to associate motif indices k_{jo} and starting times ost_{jo} to each occurrence (Eq. 16), while the document specific GEM β^o_j is used to sample the occurrence associated with each word (Eq. 17), from which generating the observations can be done as presented above (Eq. 18 to 20).

4.3 Modeling Prior H and Motif Length

In previous sections, we introduced the global structure of the models we proposed. We intentionally simplified the definition of the prior H and omitted details about it. We propose two ways of setting this prior, leading to two different models. The first model was presented in [1] was using a fixed motif length.

The model from the current article is a generalization that allows motifs to have different lengths and infers the length of each motif automatically. We named this model *Variable Length Temporal Analysis of Motif Mixtures* (VLTAMM). The model from [1], which we call TAMM, infers motifs of fixed length. This becomes a specific case of VLTAMM corresponding to a specific setting of the hyper-parameters (see Section 6). The model shown in Fig. 4 corresponds to the TAMM model that we describe first as it allows us to progressively introduce the concepts involved.

TAMM: fixed duration motifs with alignment – In the TAMM model, we use a maximum motif length that is a fixed hyper-parameter of the model. This parameter is fixed for the model and all motifs thus have the same maximum length. The parameter η (a table of the same size as a motif) defines the Dirichlet distribution prior $H = Dir(\eta)$ from which the motifs ϕ_k (defined as categorical distribution over (w, rt)) are drawn. The normalized vector $\eta' = \frac{\eta}{\|\eta\|}$ represents the expected values for the multinomial coefficients, whereas the strength $\|\eta\| = \sum_{w,rt} \eta(w, rt)$ (also noted η^W) influences the variability around this expectation. A larger weight $\|\eta^W\|$ results in lower variability.

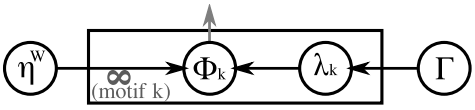


Fig. 5. VLTAMM: Variable Length TAMM model that handles different motifs' lengths. The constant temporal prior H from TAMM is replaced by a per-motif prior on length.

The parameter η' can be interpreted as the parameters of a categorical distribution. We decompose it as:

$$\eta'(w, rt) = \eta_1(w|rt)\ \eta_2(rt) \tag{21}$$

in which we define the word probabilities η_1 for a given rt to be uniform (e.g., $\forall w,\ \eta_1(w|rt) = \frac{1}{\#\mathcal{V}}$), and the η_2 prior on rt with a decreasing shape like shown in Fig. 6 (only the shape is important for now). This decreasing prior on rt plays an important role during the inference. It favors activity at the beginning of the motifs and reduces the learning of spurious co-occurrences by allowing a graceful dampening of word presence at the end of the motifs unless their co-occurrence with words appearing in the first part of the motif is strong enough. More details about the exact shape of the prior on rt and its influence are provided in appendix.

VLTAMM: variable length motifs – The TAMM model is only a restriction of the VLTAMM model. In VLTAMM, the constant temporal prior on motifs presented for TAMM, H, is replaced by a motif dependent prior as depicted in Fig. 5.

The interesting property of H when defined as a fixed-size table as in TAMM is that it is a conjugate prior of the motif distributions Φ_k. However it does not allow to vary the duration of the motifs. We needed to introduce a new distribution to match our threefold requirements: having a decreasing ramp-like shape, having a variable but finite support (keeping a time-locality for the motifs) and having a meaningful conjugate prior.

We introduce the "weight-truncated exponential" distribution. Intuitively, this is an exponential distribution that is right-truncated so that there remain only a fixed area Z under the curve (e.g. $Z = 0.33$). The obtained distribution is then renormalized to obtain a probability density function. The formal expression of the *weight-truncated exponential* distribution is given by:

$$wTruncExp^{\lambda,Z}(t) = \begin{cases} \frac{\lambda e^{-\lambda t}}{Z} & \text{if } 0 \leq t \leq \frac{-ln(1-Z)}{\lambda} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

A gamma distribution with parameters $\Gamma = (\Gamma_1, \Gamma_2)$[2] is used as a prior from which the λ parameter of each motif is drawn. It can be verified that the gamma prior is actually a conjugate of the weight-truncated exponential distribution with parameter λ and a fixed Z (detailled in appendix). This conjugacy relationship is conditioned by the fact that $L^{\lambda,Z} = \frac{-ln(1-Z)}{\lambda}$ (the length of the support) is greater than any observed draw from the weight-truncated exponential distribution. In practice, we fulfil this condition by using rejection sampling when re-sampling the λ_k parameters.

2. The α, β parameterization of the gamma distribution is used (see http://en.wikipedia.org/wiki/Gamma_distribution).

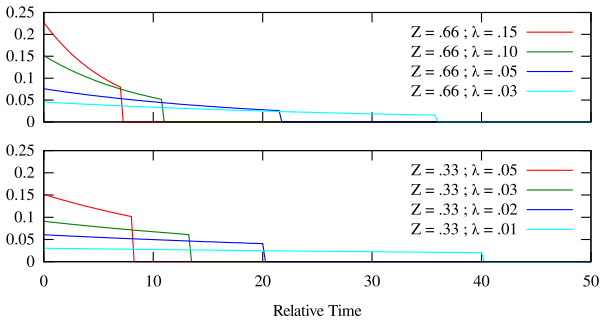


Fig. 6. Weight-truncated exponential distribution with various values for Z (truncation weight) and λ (exponential rate). This distribution family allows to control both the size of the support and the slope of the ramp.

Using the weight-truncated exponential distribution family and its gamma prior, we can model motifs of different length. We fix the Z parameter in the model and let each motif k has its own λ_k parameter. The variation of λ_k changes the effective motif length $L^{\lambda_k,Z}$. The VLTAMM model is fully defined by considering Eq. 8 to Eq. 21 and the following:

$$\forall k \qquad \lambda_k \sim Gamma(\Gamma_1, \Gamma_2) \tag{23}$$

$$\eta_2^k(rt) = wTruncExp^{\lambda_k,Z}(rt) \tag{24}$$

5 Inference

In this section, we explain the different steps involved in the inference (see appendix for more complete details). A cross-platform, standalone executable for the VLTAMM (and TAMM) model will be made publicly available.

To perform the inference, we use a collapsed Gibbs sampling scheme and sample over all $\{o_{ji}, k_{jo}, ost_{jo}, \lambda_k\}$:

- o_{ji}: association of observations to occurrences
- k_{jo}: association of occurrences to motifs
- ost_{jo}: starting times of occurrences
- λ_k: per-motif parameter

Note that given these sampled variables, other variables are deterministically defined. These are $\{st_{ji}, z_{ji}, rt_{ji}\}$.

The remaining variables, $\{\phi_k, \beta^M, \beta_j^o\}$, are analytically integrated out in the sampling. We can integrate out the motifs themselves (the Φ_k tables) as we used for H a Dirichlet distribution (which is conjugate to our categorical motif distribution). Due to space constraints, more detailed equations used in the Gibbs sampling process are provided in appendix. In the rest of this section we summarize the main elements of the Gibbs sampler for the proposed model.

Let's briefly recall the DP mixture model shown in Section 4.1 and study its relationship with the Chinese

Restaurant Process (CRP) where mixture components are drawn sequentially. For example, we can consider a DP of concentration c and base distribution H. In the CRP definition, given a set of previous draws of mixture components and data samples from the mixture components, a new draw is obtained by considering two possible cases. A new draw can either belong to one of the previously drawn mixture components, with probability proportional to the number of elements assigned to the mixture component; or, to a new mixture component drawn directly from H with a probability proportional to the concentration c. This property of the CRP, together with the interchangeability of the observations are highly used in the derivation of the Gibbs sampling equations.

Sampling $\mathbf{o}_{ji}$ for a given observation i in document j requires us to consider two cases: either to associate the observation to an existing occurrence or, to create a new occurrence and associate the observation to it.

During the sampling, the probability of associating an observation to a particular existing occurrence is proportional to two quantities. The first quantity is due to the DP/CRP on the occurrences and it depends on the number of observations that are already associated with the considered occurrence. The second quantity comes from the likelihood of the considered observation given its virtual association to the considered occurrence. From the occurrence starting time and the observation time, we can calculate the relative time rt_{ji} of the observation in the motif. Considering the prior H and all observations (across documents) associated to the occurrence motif, we can compute the likelihood of the considered observation with its relative time.

The other option is to create a new occurrence for the observation. Because of the Chinese restaurant process, it will be proportional to α. This probability of creating a new occurrence is also proportional to the likelihood of the considered observation under the hypothesis that it is associated with a new random occurrence. To process this last term, we need to consider the expected value over all possible starting times and all possible motifs for the new occurrence. With a uniform prior on the starting times, we manage to integrate over the starting times. Considering all possible motifs is more complicated: here again we have a DP and, the motif can be either an existing one (with a probability proportional to the number of occurrences across documents for this motifs) or a new motif drawn from H with a probability γ. Given our conjugate Dirichlet distribution prior H over the motifs, we manage to integrate over the new motifs drawn from H.

Sampling $\mathbf{k}_{jo}$ for a given occurrence o in a document j requires us to calculate the probability of changing the motif associated to the considered occurrence. In the same way as with individual observations, we need to consider the cases of both associating to an existing motifs and associating to a new one.

Sampling of $\mathbf{ost}_{jo}$ is handled in a grouped manner. Instead of (or in addition to) resampling each ost_{jo} independently, we group the occurrences that are currently associated to the same motif. More formally, we consider K (current number of motifs) groups: $\forall m \in 1..K,\ Gr_m = \{ost_{jo} | k_{jo} = m\}$. When we do this resampling, we are considering a common change in starting time for all occurrences of a group. To make this procedure faster, only the values $-1, 0, 1$ are considered for the time offset. Subsequent iterations of the process will make it possible to cumulate offsets in case of need. This grouped resampling speeds up motif alignment during the Gibbs sampling by making it easier to go out of non-optimal modes of the parameter distribution.

Sampling λ_k (for VLTAMM) for a given motif k is done using the conjugacy relation introduced in section 4.3. As a reminder, our prior over each λ_k is $Gamma(\Gamma_1, \Gamma_2)$. Given this prior and all observations for motif k, the conjugacy of the weight-truncated exponential distribution gives a posterior distribution $Gamma(\Gamma_1 + N, \Gamma_2 + \sum_{i=1..N} rt_i)$. In this expression, N is the number of observations associated with motif k and we sum over the relative times of all these observations. The actual conjugacy relationship holds only under the condition that the drawn λ is such that $L^{\lambda,Z}$ is greater than any rt_i. We reject any drawn λ value that does not satisfy this constraint to actually use the proper conjugacy relationship.

6 Meaning and Setting of Parameters

The proposed models have various (hyper-)parameters that can be set to influence the results of the inference. We detail the semantics of these parameters and study how we can set them. In this section, we consider the VLTAMM model, given that parameter-wise, it is a superset of the plain TAMM. As a summary of Section 4, we have the following parameters for VLTAMM: γ, α, Z, $\Gamma = (\Gamma_1, \Gamma_2)$ and η^W.

The concentration parameters γ **and** α of our Dirichlet processes affect the number of meaningful motifs and occurrences respectively in each document of our model. As presented in Section 4.1, a Dirichlet Process with concentration c represents an infinite mixture which weights are drawn from a stick breaking process. For a particular weight vector drawn from a GEM process with concentration parameter c, we can count how many of the first weights are needed to go above a fixed proportion P of the total weight. Fig. 7 shows, for various concentration c and proportion P, the probability of reaching a cumulative weight P with exactly the first n mixture components. As an *example*, we can read in Fig. 7 that with $c = 8$, most of the time, we'll need between 14 and 40 mixture components to cover $P = 95\%$ of the weight.

With the help of Fig. 7, the setting of the γ parameter can be directly translated into a prior on the number of

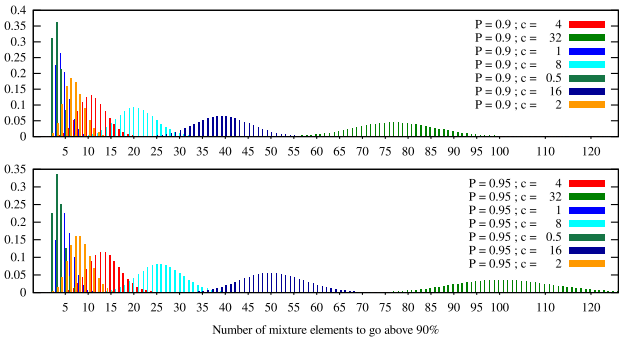


Fig. 7. Stick breaking process $GEM(c)$: distribution of the number of mixture elements sufficient to cover a proportion P of the total weight. Distributions are shown for different concentration c and proportions P (90% and 95%). (legend is shuffled to improve rendering of graphs)

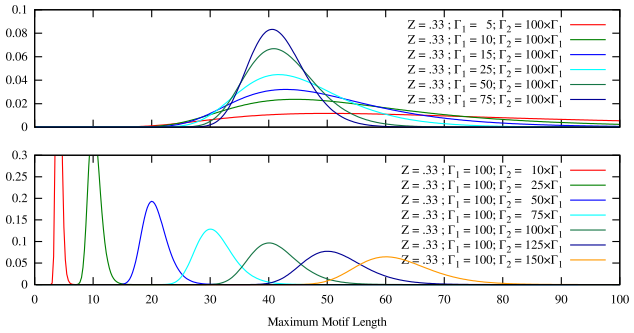


Fig. 8. Maximum motif length: distribution of maximum motif length $L^{\lambda_k,Z}$ when varying the prior Γ and keeping the parameter Z fixed. We see that Γ can be used to control the location and spread of the range of prior acceptable values for $L^{\lambda_k,Z}$.

motifs. The interpretation of the α parameter needs more attention. Since α controls the number of occurrences in a document, one might conclude that it should depend on the document duration too. In practice, by looking at the Gibbs sampling equations that are taking the data into account (see Appendix, Section 10.4), it can be shown that α is more related to the average number of overlapping occurrences on a short time frame. The consequence is that α can be set independently of the document length and that it takes relatively small values.

The fixed truncation weight Z is a shape parameter for the weight-truncated exponential family we use. It controls the shape of the temporal prior within a motif. As shown in Section 4.3 and Fig. 6, the weight-truncated exponential is decreasing on its support $[0, L^{\lambda,Z}]$. To help in choosing Z, we consider q that we define as the ratio between the value of the distribution at 0 and its value at $L^{\lambda,Z}$. From the expression of the distribution given in equation 22, we can easily derive the following relations:

$$q = \frac{1}{1-Z} \qquad Z = 1 - \frac{1}{q} \tag{25}$$

In practice, we want a reasonably decreasing ramp distribution and use $Z = 0.33$. This translates to $q = 1.5$ which means the highest point of the ramp (in 0) is 1.5 higher than its lowest point (in $L^{\lambda,Z}$) as shown in Fig. 6.

The length prior parameter $\Gamma = (\Gamma_1, \Gamma_2)$ controls the prior on λ values. Indirectly, given a fixed Z, Γ can also be considered as a prior on the length of the motifs $L^{\lambda,Z}$. To help in the selection the value of Γ, we plot the probability density function of $L^{\lambda,Z}$ for various values of $\Gamma = (\Gamma_1, \Gamma_2)$. Figure 8 shows the prior on $L^{\lambda,Z}$ for different values of Γ.

The prior Γ can restrict VLTAMM to TAMM. By using a proper Γ, we can force the motifs to have a fixed maximum length L. First, we have to compute the λ corresponding to the length L using the formula $\lambda = \frac{-ln(1-Z)}{L}$. Then we must make the gamma distribution as close as possible to a Dirac in λ.

Given our parameterization of the gamma distribution, its mean is $\frac{\Gamma_1}{\Gamma_2}$ and its variance is $\frac{\Gamma_1}{(\Gamma_2)^2}$. Thus, the gamma distribution tends to a Dirac in λ when $\Gamma_1 \to \infty$ and $\Gamma_2 = \frac{\Gamma_1}{L}$. In practice we can take a huge value for Γ_1, for example 1000 times the total number of observations in our model.

The prior weight η^W impacts the shape of the motifs as in topic models such as LDA. As well studied in [21] this parameter has a joint impact on both smoothness and sparsity of the motifs. In practice, a relatively broad range of η^W values produce meaningful results. We provide, in Section 8.2, some study of the effect of varying the η^W parameter.

7 Experiments on Video Data

In this section, we present how our model can be used on videos. Validation on synthetic data and on other datasets are presented in Section 8. A table summarizing all the parameters used to generate the figures of sections 7 and 8 is provided in appendix.

In Section 7.1, we show how to build the temporal documents from input videos. We then present in Section 7.2 the obtained results on single view traffic scenes coming from recent work on unsupervised activity analysis (MIT [3], UQM [8], far-field [11] and ETHZ [5] videos), discussing the temporal duration modeling aspects of our method. We emphasize the interest and impact of modeling time within motifs by comparing our results with other approaches (Sec. 7.3). Finally, we show in Sec. 7.4 how the method can deal with more difficult situations involving a multi-camera network of a metro station. Note that the metro context is challenging as it features less structured activities (and timing) than the above traffic datasets.

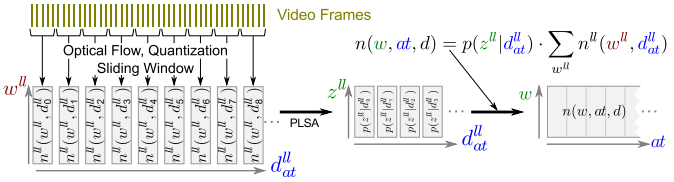


Fig. 9. From video frames to one temporal document. See text for explanations.

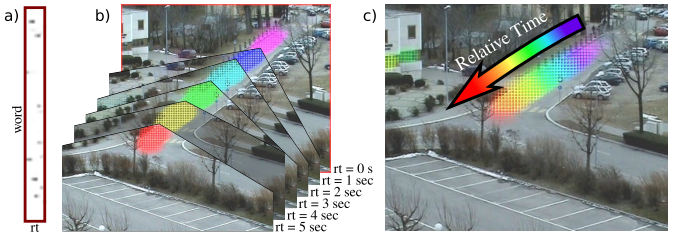


Fig. 10. Motif representations (best viewed in color) – a) motif as a matrix; b) back-projection at each relative time rt; c) using color-coded time information.

7.1 From Videos to Temporal Documents

We present here how we create temporal documents from the input videos. This means defining what is our vocabulary (what is a word) and extracting word counts at each time instant. As time step for the temporal document, we use a resolution of one second. One approach to define our vocabulary would be to directly use low-level features. However, as this results in an overly large vocabulary with high redundancy and would lead to a high inference computational load, we rely on a topic model to capture the low-level feature co-occurrences and build our high level words, as described next. To avoid confusion in the notation, we systematically use a "ll" superscript to denote the words and topics form this low level layer. Fig. 9 helps in supporting the explanation of the processing that we apply on videos.

Feature extraction – For each video, we first extract optical flow features (motion direction) on a dense image grid. We keep only pixels where some motion is detected and for these, we quantize the motion into 9 "categories": one for each of the 8 uniformly quantized directions and one for a really-slow motion. A low-level word w^{ll} is then defined by a position in the image and a motion category. The size of this low level vocabulary is rather large initially but can usually be reduced to around 30000 words considering only words that are actually observed. We then run a sliding window of 1 second long (5 to 30 frames, depending on the dataset) without overlap and collect, at each time instant, at (absolute time) a histogram $n^{ll}(w^{ll}, d^{ll}_{at})$ of the low-level words in the corresponding time window. Here, d^{ll}_{at} denotes the low level document obtained from the sliding window at a time at.

Details on dimensionality reduction – On the (unordered) set of documents $\{d^{ll}_{at}\}_{at}$, we apply the Probabilistic Latent Semantic Analysis (PLSA) algorithm. PLSA takes as input the word counts $n^{ll}(w^{ll}, d^{ll}_{at})$ for all documents d^{ll}_{at} and learns a set of topics, where each topic z^{ll} is defined by a distribution $p(w^{ll}|z^{ll})$ over the low-level words and corresponds to a soft cluster of words that regularly co-occur in documents. By using these topics as our high-level words we obtain a scene-adapted vocabulary while implicitly achieving a dimensionality reduction.

PLSA, both during its learning phase and its application to new video documents (i.e., keeping the low-level topic fixed) also produces a decomposition of each document d^{ll}_{at} as a mixture of the existing topics where the topic weights are given by the distribution $p(z^{ll}|d^{ll}_{at})$. We use this information, re-weighted by the amount of activity (represented by the number of observed features in d^{ll}_{at}, $n^{ll}(d^{ll}_{at}) = \sum_{w^{ll}} n^{ll}(w^{ll}, d^{ll}_{at})$), to build the temporal documents which constitute the input of our model. More precisely, the (high-level) word counts $n(w, at, d)$ at each time instant at in our temporal document d is expressed as:

$$n(w = z^{ll}, at, d) = p(z^{ll}|d^{ll}_{at}) n^{ll}(d^{ll}_{at}). \quad (26)$$

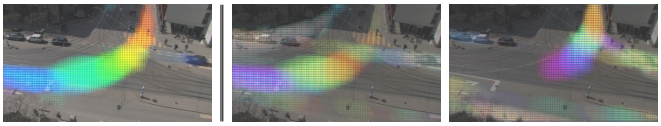
Note that in this paper, we use PLSA to find our local scene topic for dimensionality reduction. Latent Dirichlet Allocation (LDA) or its non-parametric extension using Hierarchical Dirichlet Processes (HDP-LDA) can be directly used in the same way as we did in [22]. As we operate a conservative soft clustering, the exact technique used has little influence on the global results.

7.2 Recovered Motifs

Motif representations – We apply our model on different video datasets to retrieve recurrent activities as motifs. A recovered motif is a table providing the probability that a word occur at a relative time instant with respect to the beginning of the motif as exemplified in Fig. 10a). Since, as introduced in Section 7.1, each word w corresponds to the response of a PLSA topic z^{ll}, we can back-project the set of locations where it is active in the image plane[3]. Subsequently, to visualize the content of each motif, the word probabilities for each relative time instants rt are back-projected into the image plane to obtain activity images I_{rt} as shown in Fig. 10b).

From the back-projected images, a short video clip can be generated for each motif. To show examples in this paper, we use a static, color coded representation illustrated in Fig. 10c). Each time instant is assigned a color (from blue/violet to red) and superimposed in a single image. This representation is more compact than

3. Note that the PLSA topics contain more than the image location: their low-level word distribution $p(w^{ll}|z^{ll})$ provide information about the motion direction distribution as well. However, for purposes, only the location probabilities obtained through marginalizing the motion dimensions are displayed.



10 second long 5 second long

Fig. 11. Kuettel-2 (ETHZ) dataset. Example of a motif split into two motifs when the maximum allowed duration is below the real activity duration. (colors: see Fig. 10)

showing all images but suffers from occlusions when motion is slow (e.g., blue is occluded by cyan and green).

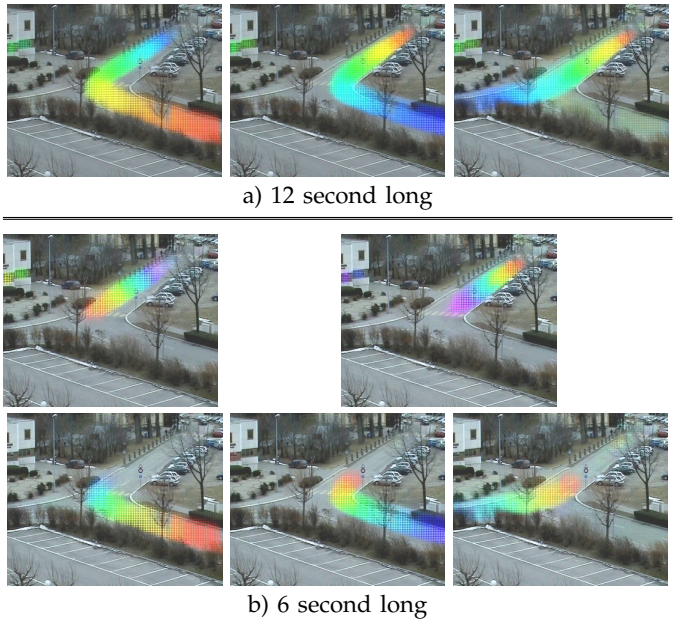
Even with such a compact color-coded representation, the amount of results that can be provided in the body of this article is limited. Generally, The observed results are interesting and the motifs recovered by our method actually correspond to real activities. Here, we provide sufficient illustrations to underline the key behavior of our model that came out of the analysis of the results.

Impact of the motif duration in TAMM – As explained in Section 4.3, the model allows both a soft and hard setting of the motif maximum duration through a prior. To see the benefit of each possibility, we can first explore the advantages and limitations of using a hard prior on the motif duration (i.e., using a fixed duration as in our TAMM model, or as in [1] and [11]).

When the hard maximum allowed motif duration is shorter than the real duration of an activity, this activity is usually cut into multiple motifs (the different parts can not get fused). This phenomenon is illustrated in Fig. 11 and Fig. 12. The right part of Fig. 12 also illustrates another typical situation where two long motifs share a common subpart which is well factored as a single motif when the maximum allowed duration is lower than the actula full activity duration. Finally, if we increase the motif duration beyond the actual activity duration, the model usually recovers the same motifs properly.

TAMM (fixed length motifs) have been shown to behave well when changing the maximum duration. However, setting a hard limit on motif duration has also some potential problems. An activity whose duration is just beyond the maximum might be split or have a small portion at the beginning or end removed (and potentially merged with another motif). At the other end, augmenting the maximum duration increases the chance of capturing spurious activity co-occurrence, especially when the training data is small and some activities are much shorter than the set maximum duration.

Through the use of a (soft) prior on the motif duration parameter, VLTAMM provides a principled approach to deal with the fixed-motif duration shortcomings and activities of different durations within the same scene. In the above examples, VLTAMM allows us to roughly set the prior on length (e.g., around 10 seconds) and recover the real activity duration. More illustrations of this ability are provided later (Sections 7 and 8).



a) 12 second long

b) 6 second long

Fig. 12. Far-field dataset. Example of motif split when the maximum allowed duration is below the real activity duration. The motifs on the right in a) have a common subpart. As a result, these 2 motifs get split into only 3 shorter motifs on the right of b). (colors: see Fig. 10)

Variable Length Motifs with VLTAMM – We experimented with VLTAMM (allowing variable length for motifs) and show here some results for the UQM junction dataset in Fig. 13 (detailed below). Due to traffic lights, the scene undergoes cyclic behavior with a period of around 80 seconds.

We set the length prior of VLTAMM to $(\Gamma_1, \Gamma_2) = (50, 2500)$. This prior, as can be derived from Fig. 8, Section 6, is relatively broad with an average length of 20 seconds. Results are shown on the left part of Fig. 13. VLTAMM recovered exactly 4 motifs having lengths ranging from 18 to 23 seconds and covering the full cycle and aligned to the instant where no or less activity is going on in the cycle.

Long Motifs – Our VLTAMM model is also capable of capturing long activities. The right part of Fig. 13 shows the results obtained for "UQM junction" scene introduced before. This time, we set the motif length prior of VLTAMM to $(\Gamma_1, \Gamma_2) = (100, 20000)$. This corresponds to a smooth prior centered around 80 seconds.

In this setup, VLTAMM automatically recovers a single major motif explaining 75% to 99% of the observations depending on the runs. This motif captures the full traffic cycle with the successions of phases. Due to the activity duration and large amount of spatial overlap during these phases, the motif static color-coded representation shown in Fig. 13 is highly obfuscated. Overall, the recovered motif is really close to the concatenation of the 4 motifs obtained on the left part of the figure.

This example shows that the proposed method can

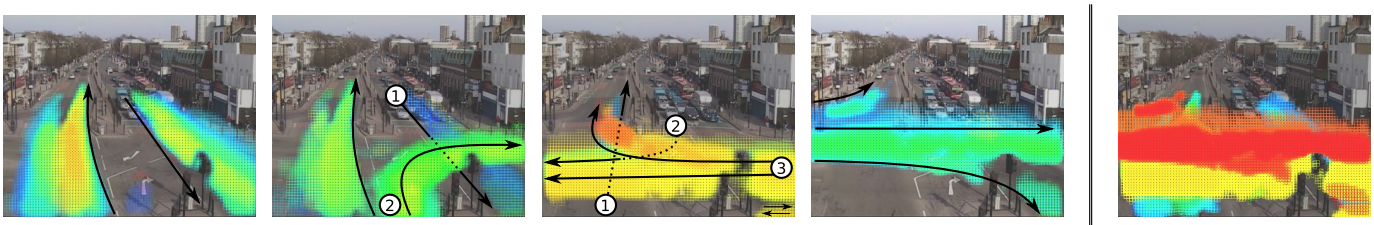


Fig. 13. UQM Junction dataset. Comparison of short and long motifs obtained with two different settings for the VLTAMM prior on motif length. Motifs are representing more than 95% of the observations. (colors: see Fig. 10)

capture scene-level temporal motifs in the same way global Hidden Markov Model based methods [8] would do, but extracting in addition a more detailed representation of the spatio-temporal content.

Additional results – Due to the amount of space required to show the results for a scene, we limit the amount of results provided in the body of the article. In addition to the previously covered results, we provide in Fig. 14 and Fig. 15 the major motifs for two other scenes. These scenes are taken from [5] and we used respectively 43 and 86 minutes of video to learn the motifs.

In Fig. 14, we see that the motifs capture the different activities of cars: depending on where they come from the cars have different speed and typical trajectories, as explained in the caption. The tram lanes in both directions are also captured. We also observe that some motifs capture interactions between object. For example after car passing, the pedestrian starts crossing the zebras (motif of the lower row, second column).

In Fig. 15, the motifs capture all possible car and tram lanes. Some motifs capture mixes of trajectories that tend to co-occur in the training set due to the simultaneous start after some traffic light changes. This includes car flows in both directions, car flow splitting in two, and simultaneous car and tram motion.

Activity diary and abnormality reporting Our models captures meaningful temporal patterns and is able to find when they occur. We can easily take a video (e.g., 30 min or 1 hour long) and measure the amount of occurrence of each motif trough time. We also extract an abnormality measure based on reconstruction error as in [22]. These source of information together with snapshots of the most abnormal instant are automatically consolidates in an activity diary. For space reasons, we provide such diary in appendix. In the case of the far-field dataset, the most abnormal instants are unusual car trajectories or unusual car speed, often due to big trucks altering the traffic.

7.3 Comparison with other methods

To illustrate and validate the advantages of our approach, a comparison with other approaches has been carried out. Some methods such as [5] have high difficulties to extract local activities with temporal information properly. and only model global, scene-level activities as in [8]. We consider the methods used in [3] and [9] as baselines to qualitatively compare the topics learned from them with the motifs learned from our method.

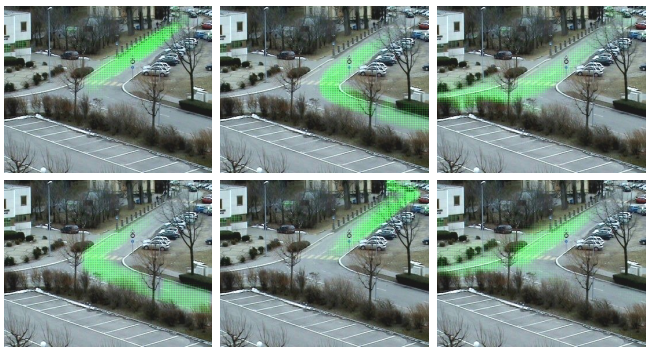


Fig. 16. Topics of method from [3]: LDA variants applied on a sliding window, ignoring time information. Resulting *patterns capture no temporal information*.

The first kind of methods, used by [3], features a sliding window scheme, where low-level features are gathered over temporal windows to build independent documents in which the temporal information is ignored. On these documents, a topic model like LDA, HDP-LDA as in [3] or PLSA is applied. This corresponds to our low-level processing method (cf Section 7.1), but using a longer temporal window (we use 10 second windows in the example below). As the method explicitly ignores time ordering, the recovered motifs, although being globally meaningful and relevant, do not carry any temporal information as shown in Fig. 16 (where motifs are all green). This absence of time information within topics results in a temporal granularity loss when trying to localize the advent of these activities compared to our scheme. This is illustrated in Fig. 17: while our method provides clear starting times (although not perfect), all others methods show noisy or very smoothed topic response. Here, the words capture a lot of information and thus a plain LDA exhibits reasonable but noisy response. With less informative words, our method provide provides an even greater gain, as shown in a counting task in Section 8.1.

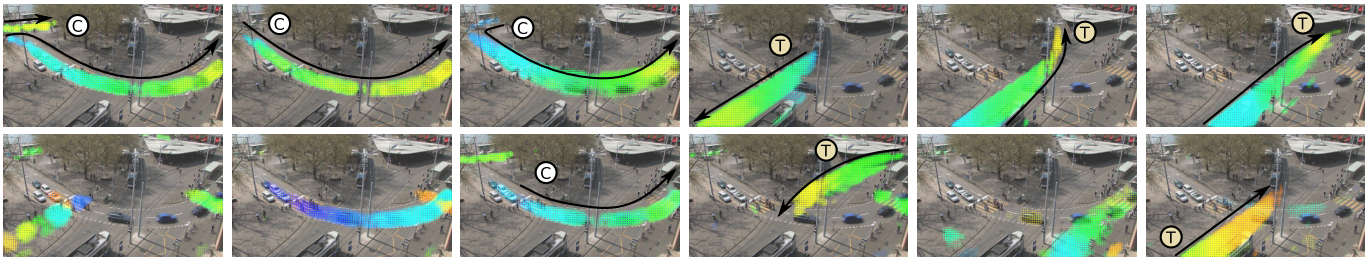


Fig. 14. Kuettel-4 (ETHZ) dataset: 12 most probable motifs representing 95% of the observations. We annotated with a C the car motifs (and with a T for tram motifs) that might look the same at the first glance. To help in seeing the differences, the curved black arrow has been put at the same location for the four motifs. They correspond to 4 different activities (from left to right): 1) cars arriving from the left of the image and turning to their right to the visible road; 2) cars driving straight into the road visible in the video; 3) cars arriving from the top and turning to their left, we see that they use more the left lane of the visible road; 4) cars starting to move when the traffic light (in the middle of the image) changes to green. All T motifs are for different tram lanes. The last T motif is actually used to explain a tram that stopped because of some pedestrians (the motifs also gets some support with slower trams).

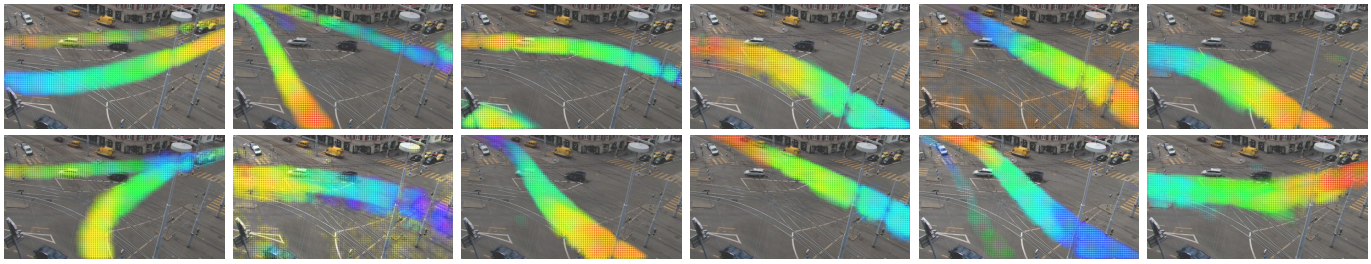


Fig. 15. Kuettel-3 (ETHZ) dataset: 12 most probable motifs representing 93% of the observations. All car and tram lanes are captured as motifs. Two motifs with notable weight are not shown but capture additional lanes. There are a lot of different lanes (and tram tracks) in this scene and no redundant motifs are obtained. (colors: see Fig. 10)

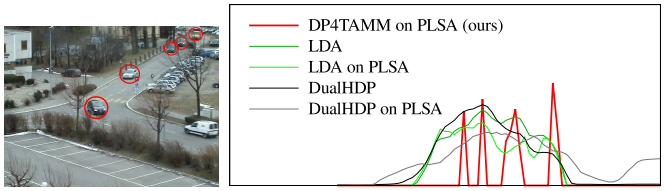


Fig. 17. Comparison of the amount of presence of selected topic/motif (corresponding to the considered car activity, such as the first image from Fig. 16) against time using different methods in a case where 5 cars follow each other. Our method shows distinctive peaks while all others are confused by the loss of temporal information.

The second method, called TOS-LDA (Time Ordered Sentive LDA) [9] uses the same sliding window scheme and LDA topic model with a modification. Unlike in the previous approach, the temporal information within the window is kept, i.e., the vocabulary used for topic modeling is actually the Cartesian product of the original feature vocabulary and the relative time within the window. The issue with this model is that all documents are considered independent and are not necessarily aligned with the actual start of the activities in the video. As a result, a real activity gets cut in different places by the sliding window process and the topic modeling has to model all the possible different starts of activities in their topics. The result of the method on the Far-Field data is given in Fig. 18 and clearly exemplify this phenomenon.

Our approach does not have the drawbacks of the above two kinds of methods. It captures the temporal information within activities while also aligning them with their start times automatically. By aligning the itself with the real starting of the activities, our model is able to both capture temporally meaningful patterns and avoid noise learning issue.

7.4 Calibration-free multi-camera analysis

We also explored the application of our method to capture temporal dependencies among multiple cameras. To this end, we considered cameras from a metro station and report results with 4 cameras in this paper. Note that such an environment is much more challenging than in the urban case given that activities are less structured, both spatially and temporally. The recordings are made at 5 frames per second, with an image resolution of 352x288. We used 2 hours of video captured on a random day from 7:00AM to 9:00AM. Fig. 19 shows the locations of the cameras (numbered from 1 to 4) on a schematic map, while Fig. 20 shows the camera views. As can be

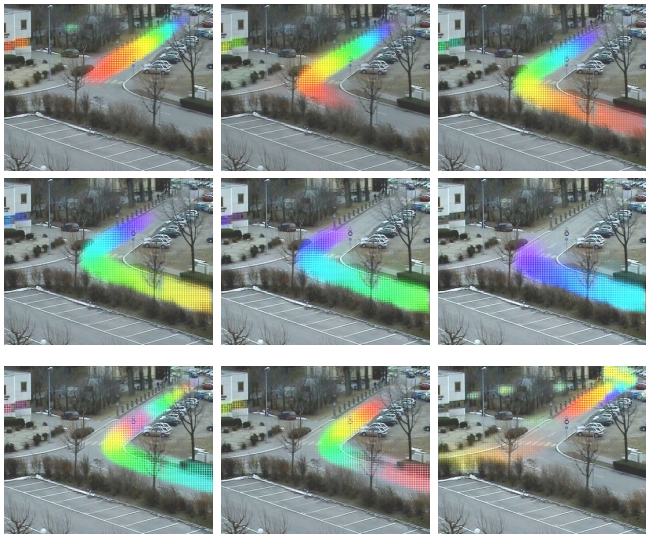


Fig. 18. TOS-LDA [9]: top 9 recovered motifs out of the 15 using a window of 12 seconds. The motifs capture temporal information but *multiple redundant motifs are necessary* to cover all possible "cuts" of an activity (first 2 rows). Some motifs also mix up unrelated activities (lower row), as illustrated by the presence of the same color at several places. (colors: see Fig. 10)

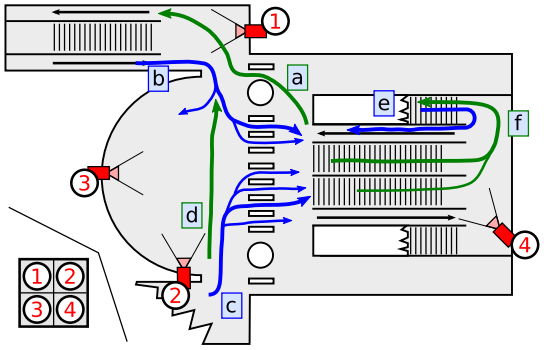


Fig. 19. Schematic map showing the location of the 4 (numbered) cameras used for the calibration-free multi-camera analysis experiments. *L*ower left corner: layout of the camera views used to display the captured motifs in Fig. 20. *L*etters and Arrows: summary of the activities captured by the motifs displayed in Fig. 20..

seen, some field of views have a notable overlap (e.g., those of cameras 2 and 3) while others are just distantly connected.

To process the multiple cameras, we adopted an early integration scheme. We jointly processed the low-level features through our method by creating a low-level count matrix $n^{ll}(w^{ll}, d^{ll}_{at})$ as the concatenation of the count matrices of all cameras. In other words, we used as vocabulary the union of the vocabulary from all cameras. Once the low-level count matrix is obtained, the processing is exactly the same as in the single camera setup. Note that this is equivalent[4] to the processing by our approach of a single video created by sticking together at each time instant the frames of all video streams (cf concatenated views shown in Fig. 20).

Automatic low-level camera soft calibration – Given the variety of possible instantaneous motions that can happen in the 4 views, we used 300 low-level PLSA topics (see Section 7.1). Due to the huge amount of space that would be required to show these 300 topics, only higher level motifs are shown in this article. Still, already at the low level, the behavior of our approach is interesting. Although the 4 views are processed jointly, most of the low level topics have their support in a single camera view. In the case of view overlap, as is the case of cameras 2 and 3, we also obtain low level topics that span these two views. A few noisy topics also capture some random co-occurrences (within a view or across views): this behavior would likely be solved if a larger amount of training data was used. In summary, this demonstrates the ability of the low-level co-occurrence analysis to capture relevant inter-camera relations, thus achieving a soft calibration task automatically.

Recovered motifs from multi-camera – For space reasons, Fig. 20 shows only the 6 most probable motifs recovered by our algorithm when setting the prior on the average maximum duration to 30 seconds with the VLTAMM model. For comparison, topics recovered by LDA are shown in appendix. These motifs represent 68% of the observations. They are named with letters from a) to f) and the activities they capture are schematically represented in Fig. 19. Despite the complexity of people behaviors and trajectories, our method properly finds the relations between cameras without any calibration or supervision. We obtain motifs covering up to three cameras like motif a) and b) or two cameras like motifs c) and d). For instance, the motif in Fig. 20a) corresponds to people on the left of view 3 (after arriving from the escalator or from the left corridor), disappearing behind the pillar for some seconds, then reappearing in view 2, and finally exiting the metro using the right escalator in view 1.

Timing information evaluation – Since motifs comprise temporal information, a question that arises is whether the timing it models matches the actual timing of the real activities it captures. This problem is particularly relevant for motifs that spans multiple cameras. To evaluate this aspect, we have selected 3 motifs from Fig. 20 that recover typical paths performed by people accross at least two camera views. The path start and end locations were defined from the motif backprojection (see Fig. 10) and the duration was obtained from the time difference between the corresponding instants within the motifs. As ground truth, for each path, we measured on 10 to 20 minutes of video (depending on path frequency) the

4. Except for spurious differences in motion extraction at boundaries.

TABLE 1
Typical path durations (in seconds) as measured from the video or as recovered from the motif. Path are identified by their Start and End location shown on the corresponding motif in Fig. 20.

Path		Measured duration					Duration
motif	start-end	avge	std	min	max	med	(motif)
a)	A-B	24.4	2.4	20	28	24	26
b)	A-B	24.9	4.8	20	35	23	24
b)	A-C	18.6	8.3	11	35	15	17
d)	A-B	7.5	0.9	6	10	7.5	8

time taken by people to travel from the start to the end of these paths (not necessarily taking the same trajectories), and report the obtained statistics.

Results are shown in Table 1. As can be seen, in general the timing provided by the motifs are very close to the average or median durations measured from real trajectories. We can notice the larger variation in entering the metro (second path, from A to B in motif b) as compared to when exiting the metro. This is due to the fact that entering people use different turnstiles, can queue for some seconds, or are taking more time to hand out their pass/ticket. Similarly, large variations are observed in the third case, as people are sometime looking at the poster/map on the wall before reaching the path end point. However, as these are less frequent, the median value is smaller in this case. The discrepancy with the reported motif duration in the table is mainly due in the uncertainty in identifying the path end time from the backprojection: the motif exhibits a mass of activity (probability) in front of the poster from 13s to 17s (yellowish region on the left in view 2), reflecting well the inherent variability of the data, but we have selected 17s as the end path as the backprojection at that instant seemed more localized in the image.

Difficulties and amount of training data – The motifs from Fig. 20 also illustrate some general difficulties encountered when dealing with such challenging dataset featuring dense activity. Continuous motion activity from the rolling escalator in camera 1 are present in all motifs; this is not a problem in itself but we would have liked to see it separated from the rest. Secondly, some fortuitous co-occurrences are captured as visible especially in motif f). Given the size and complexity of the state-space (low-level vocabulary of around 20,000 words, 300 high-level words, average motif duration around 30 time steps (seconds)), variations in activities, and the low amount (duration) of training data in comparison, the results still demonstrate the model's ability to capture meaningful temporal patterns.

The multi-camera results were obtained using two hours of unlabelled videos. We also experimented with only the first hour of the same dataset which contains 40% of the observations of the two hour video. In this case, the algorithm recovers comparable motifs but the presence of spurious co-occurrences is exacerbated. These results along with others suggest that we could obtain cleaner motifs by using a larger amount of training data. Alternatively, if the amount of co-occurrence becomes too large (eg increasing the number of cameras or having crowded scenes), the use of tracklet in place of optical flow as information unit could decrease the number of spurious matches, but would require another modeling approach [23].

Fig. 21. Sample temporal document extracted from microphone array signals [20]. This document features 4 cars coming from right to left and one from left to right.

8 Experiments on other data types

Our model is not limited to video data. Other kind of features or data types can be used as well, taking even more advantage of temporal modelling. This section consider the case where features contain no temporal information (contrary to optical flow).

8.1 Car counting using an audio microphone pair

We apply our model on an audio dataset provided by EPFL [20]. This dataset is recorded by a microphone array (two microphones separated by 20 centimeters) placed on the side of a two way road. In [20], detecting (and counting) vehicle is one of the two addressed tasks. While the authors of [20] use a dedicated algorithm, we can use our model for this task, as it is able to automatically discover recurrent events and when they occur. More precisely, we can create a detector by simply thresholding the motif occurrence weights and obtain the set of instants when the motifs occur.

From audio signals to temporal documents – Our model requires the definition of words. In this audio case, our words w_θ are defined by the audio activity coming from a direction $\theta \in [-90, 90]$, where a 0 direction corresponds to a vehicle being in front of the microphone array. The observation of words at a given time instant at is obtained by computing the Generalized Cross-Correlation (GCC) between the microphone signals at this time, which provides a measure of the sound intensity in the Direction Of Arrival (DOA) θ. More precisely, at each time instant, the set GCC measurements are first normalized and a uniform distribution is subtracted from the resulting values. The normalization step provides some invariance to car loudness (that depends on the distance to the microphone), while the subtraction step removes uniform noise that might have been amplified by the normalization step.In practice, we used 25 words (DOA values of θ) and each time step covers 82ms. An example of resulting temporal document is provided in Fig. 21.

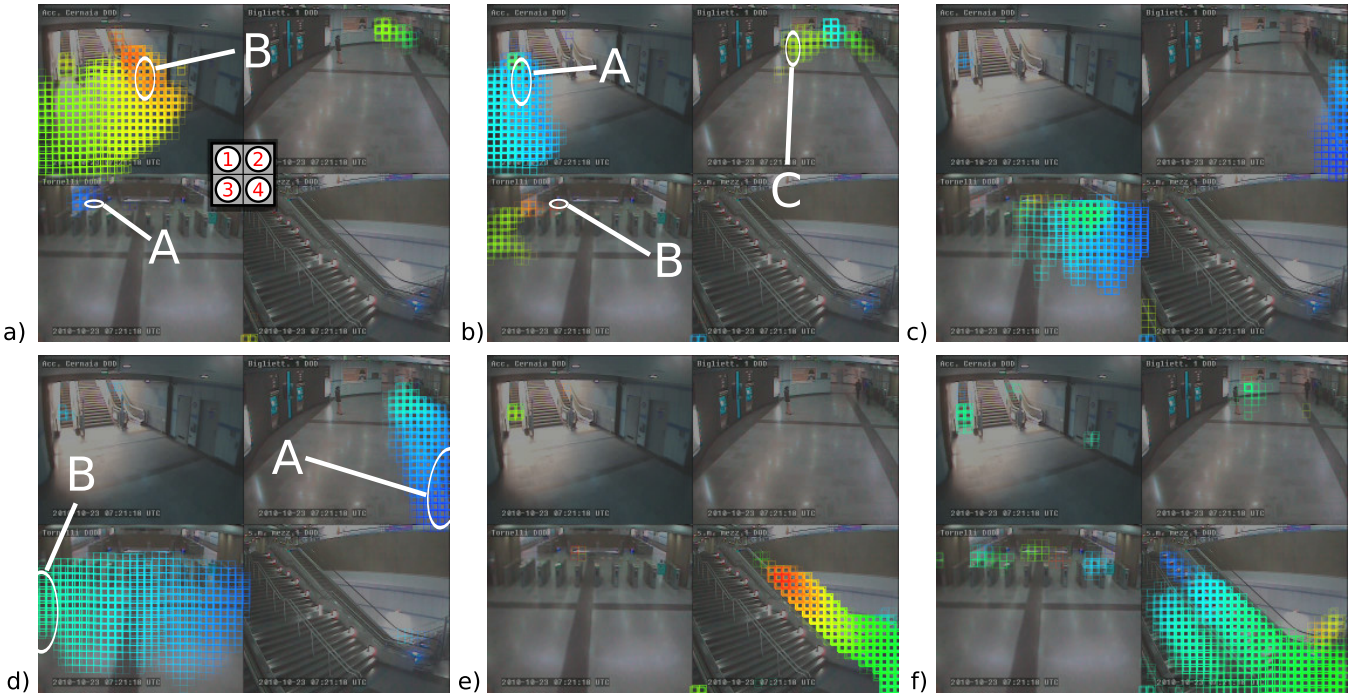


Fig. 20. Visualization of the multi-camera results. The 6 most probable motifs are shown, and the activities they capture are represented in the metro map of Fig. 19. The overlap between cameras (e.g., cameras 2 and 3) is properly found as in motif d) where people are crossing the station and also in motifs b) and c). The link between different camera field of views are also captured in motif a) where people are leaving the station, being successively visible in cameras 3 then 2 then 1. Motif b) shows the same kind of trajectory but in the opposite direction. Motif e) and f) are mainly on camera 4. Motif f) corresponds to people going down the visible stairs, then taking some other stairs down to the platform (in yellow). Motif e) shows the opposite direction, where people tend to take the escalator to go up. (colors: see Fig. 10)

Recovered motifs – Fig. 22 shows the two typical sets of motifs we obtain if we run our algorithm repeatedly with different maximum motif length priors. The effect of the motif length on the motifs is not perceivable until it is sufficiently long. From a prior maximum duration of 40 up to at least 100, the model provides the same results as in Fig. 22a). The case of Fig. 22b) has been obtained with a prior maximum duration of 20 time steps and we get similar results up to 40 time steps.

All recovered motifs are shown in Fig. 22. In these settings, there is a clear cut difference between interpretable or useful motifs and motifs that are just side products of our inference. This can be seen clearly by the weights of the recovered motifs: in both settings, the last motif represents less than 1% of the observations, the other ones ranging from 8% to 35%.

Car counting evaluation – There are two main events of interest in this setup: a vehicle going from left to right (L→R) and a vehicle moving from right to left (R→L). As presented in Section 4.2, the model inference generates a mixture of occurrences for each document: each occurrence has a starting time, a motif and a weight (proportional to the number of observations associated to the occurrence). As our method is unsupervised, to each of the two event types we would like to detect, we manually associate its corresponding discovered motif (motifs with thick borders in Fig. 22). Then, by thresholding the occurrence weights of these motifs, the occurrences can be interpreted as detections. When the value of the threshold is varied, different precision and recall compromises can be achieved. A temporal alignment is also needed as the annotated events may correspond to various positions with respect to the captured motif: we automatically tried different constant offsets for the evaluation.

To perform the evaluation, we consider 30 clips of 20s each: they add up to 10 minutes and 7320 time steps. The ground truth contains 57 event of cars going from left to

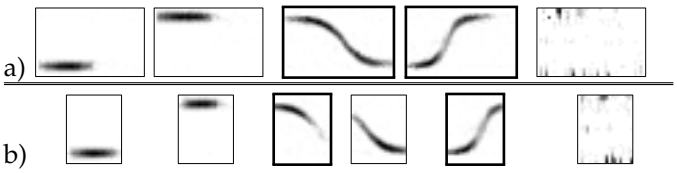


Fig. 22. Example of two typical sets of motifs recovered by our model from documents like the one in Fig. 21. a) using a prior maximum length of 60. b) using a prior maximum length of 20 (the motifs were reordered to match those of a)). In both cases, the last motif represents less than 1% of the observations.

right and 47 from right to left. We do a proper matching of the response from the detector with the ground truth: we allow each detected event to be matched with only a single ground truth event and vice versa. We accept up to 2 seconds between a detected event and the ground truth to consider a match.

Fig. 23 shows the interpolated precision/recall curves we obtain with our model, along with the operating points provided in [20]. For our model, we provide the precision/recall curves obtained by selecting the relevant event motifs from one or the other of the two sets of motifs shown in Fig. 22.

Comparing our results with those of [20], we observe that, despite being general and not tuned to the dataset, our method has similar performance as the domain specific approach from [20]. We observe as well that the performance for detecting cars coming from the right is lower than that for detecting cars coming from the left. This is due to the former ones being occluded by the later ones during crossing.

We also observe that for right to left events, the shorter motif from Fig. 22b) produces better detection accuracy. Looking at the occurrences of the longer motif, we see that there are sometimes multiple occurrences for a single car. This phenomenon is due to the variations in car speed and size that create traces of different thickness. To explain a thicker trace in full, the model needs to create two occurrences (of the long motif) starting close of each other. Some post-processing of the occurrence could improve the results: neighboring occurrence of the same motif could be merged to create a more precise detector. We haven't explored this solution further.

Comparison with other Topic models Counting results obtained with LDA and TOS-LDA are also shown in Fig. 23. Fig. 23 clearly illustrates the advantage of our method: a non-maxima suppression procedure has to be applied for other methods and their performance is lower. The reason is the following: after losing temporal information, all models like LDA, HDP and DualHDP cannot differentiate between L→R and R→L events (the order of the word ramp is what matters).

Fig. 24 illustrates this reason, showing the amount of some selected motifs/topics across time for two temporal documents. Our methods exhibits clear peaks for one motif at each "ramp up" pattern. With LDA, an higher layer would be needed to capture the succession of topic: here, a "ramp up" is the succession topic #1 - topic #0 (and a ramp down is the contrary). Capturing this temporal succession of events is exactly what our model is designed to achieve, here directly at the word level but at the topic level for video data. TOS-LDA gets confused when cars are following each others.

8.2 Parameter exploration on synthetic documents

Synthetic Dataset – To validate our model and its implementation, we apply it on synthetic temporal documents.

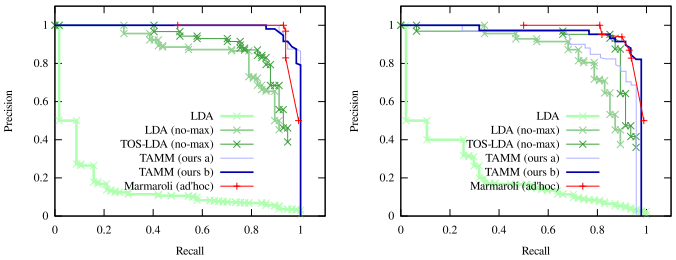


Fig. 23. Precision/Recall curves for event detection for the two kinds of events. Left graph: L→R, cars going from left to right. Right graph: R→L, cars from right to left.

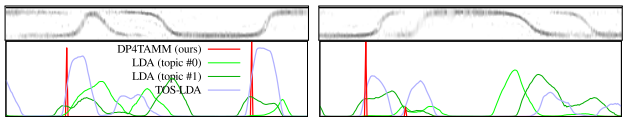


Fig. 24. Two temporal documents and the topic/motif presence using different methods (see article body).

We create motifs by rasterizing random strings made only of lower case letters. Example are given in Fig. 25.

Fig. 25 illustrates some properties that can be observed in real life scenarios and that are present in our synthetic datasets. First, the motifs content is mostly concentrated on some words (the ones from the middle as most lower case letters are small and close to the text baseline). Second, there is a notable variation in motif length (e.g., "oji" is notably shorter in Fig. 25). Last, motifs can share common subparts (e.g., letters like "g").

Each synthetic dataset has been generated as follows. First, following some rules on their number and length, we create a set of motifs. Then, using these motifs, 10 temporal documents are generated, each document being 300 time instants long and containing 5 random occurrences of motifs (see Fig. 26). Most of the obtained temporal documents contain a lot of overlap between occurrences.

Estimating the right number of motifs – We want to quantify how well our model estimates the number of motifs present in a dataset. We generated different datasets containing between 1 and 5 motifs. Our model is then used to discover the motifs and their number. Provided results are statistics over 10 runs.

We fixed the maximum length prior to 100 time steps which is largely sufficient for the motifs we consider (for which the actual maximum length is below 50). We used $\eta^W = 50, \gamma = 5, \alpha = 0.01$ for this experiment.

Most of the times, the number of motifs is exactly found and the motifs are properly recovered. For extreme values of the parameters (see analysis below) the algorithm may produce results with less motifs (a pair of actual motifs are merged into a single recovered motif) or more (an actual motif is split into two recovered motifs).

Due to the stick breaking process and the model inference (sampling process), the method might output some low-weight motifs that are not very meaningful.

Fig. 25. Examples of 6 synthetic motifs. Horizontal axis: time. Vertical axis: word features as in Fig. 2.

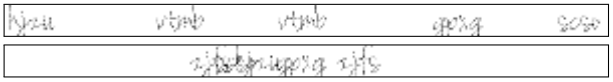


Fig. 26. Example of 2 temporal documents coming from the same set of motifs. Each document has 5 occurrences but happen to have different level of overlap.

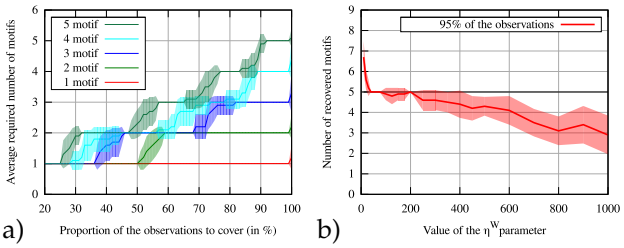


Fig. 27. Number of recovered motifs sufficient to explain a given proportion pr of the observations. Standard deviation is represented using shaded area. a) estimated number of motifs in function of the proportion pr, for different number (1 to 5) of motifs in the dataset. Note that when the threshold $pr = 95\%$, an accurate estimate of the number of motifs is found by our model. b) estimated number of motifs for different values of the weight η^W of the motif prior, and for 5 motifs in the ground truth.

Thus, in practice, we define as "estimated number of motifs" the lowest number of them which is sufficient to explain a percentage of the observations.

Fig. 27a) shows an analysis done on number of motifs recovered for datasets containing 1 to 5 motifs. The graph plots the number of motifs used by the model (y axis) to explain a percentage of the data (x axis).

While considering the curve for 5 (equiprobable) motifs in the ground truth, we observe 5 plateaux: one with value 1 around 20% (1 motif was found sufficient to explain 20% of the observations), one at 40% and so on. Other curves exhibit the same plateaux behavior.

Finally, we can observe in Fig. 27a) that any threshold value between 91% and 99% leads to proper estimate of the actual number of motifs. Note that for threshold values very close to 100%, there are alway some extra motifs that are randomly 'tested' as part of the stick breaking process, so that the number of motifs is above the actual one and the variance increases.

Effect of the η^W parameter – The η^W parameter controls the strength of the prior on motif distributions. A larger η^W should thus lead to lesser motifs that are smoother. However, η^W also plays against the concentration parameters γ (which acts as a prior on the number of motifs) and the actual information from the data.

We experimented with the same setup as above with exactly 5 motifs in the documents, and studied the estimated number of motifs while varying η^W. The concentrations parameters have been kept fixed as above ($\gamma = 5, \alpha = .01$). Fig. 27b) summarizes the behavior we observe.When η^W is too small (here, below 20), the activities are over-segmented and too many motifs are created by the algorithm. When η^W is too big (here, above 200), both the number and the quality of recovered motifs decrease: some motifs start to be a mix of several real motifs. Most importantly, a broad range of η^W values (30 to 200) produce good results.

Effect of concentration parameters and accuracy of duration estimation – As introduced in Section 4.2, our model have two Dirichlet processes. For space reasons, the study of the effect of these parameters, together with the study of recovered motif duration, are available in appendix. Experiments show that the model is stable with a broad range of concentrations and estimates well the duration of motifs.

9 Conclusions

This paper introduced a new topic model capable of automatically finding recurrent temporal patterns (motifs) in time series. The model automatically and jointly finds: 1) the number and form of motifs common to a set of temporal documents; 2) how many times and when these motifs appear in each documents; and 3) an estimate of each motif duration.

The model has been validated on a wide range of dataset including videos, signals from a microphone pair and synthetic documents. On video data, the proposed model is able to isolate and extract motifs that correspond to activities such as vehicular movement or typical person trajectories. Applied simultaneously on multiple cameras from a network, the method finds the possible links between camera views properly without using any calibration information.

On audio signals coming from two microphones, our unsupervised method has been shown to properly recover the activities of interest and yields event detection precision and recall equivalent to a more domain specific method. Using synthetically generated data, we assessed the robustness of the model to variations in its hyper-parameters. Overall, our model has been shown to produce meaningful results in various situations and could be applied to an even wider range of time series. The design of our model makes it most suitable for cases where the observed time series are the superposition of multiple unsynchronized activities.

Future work – For the evolution of the model, we foresee room for improvement at different levels. First, our model finds recurrent activities but is not designed to handle the global succession of activities. For instance, in the case of traffic cycles, we can use hierarchical approach as in [24] to capture the repetition of the

global scene cycle itself. Second, within the model, the execution speed of activities could be integrated: instead of recovering two motifs for different execution speeds, we would recover a single motif and some speed information. Last, our motifs are just probability tables (multinomial distributions) and we could either use some more constrained parametric forms or add some side information from the videos that constrains the motif learning at consecutive time steps.

The model can also be applied to new domains, data types or even features: in dense scenes like "junction", motion co-occurrence is insufficient to isolate individual cars and the use of tracklets as in [23] can prove a valid alternative to optical flow Performances of the inference on big datasets (e.g., learning on the 4 camera setup took around 10 hours on a standard PC) might also get critical to scale to broader setups. Further hierarchical structures, stochastic optimization, variational inference, parallelization and use of GPUs are other directions.

Acknowledgments

The authors acknowledge the Swiss National Science Foundation (Project: FNS-198, HAI) and from the 7th FP of the European Union project VANAHEIM (248907).

References

[1] R. Emonet, J. Varadarajan, and J. Odobez, "Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model," in *CVPR*, 2011. 1, 3, 5, 10

[2] Y. Yang, J. Liu, and M. Shah, "Video scene understanding using multi-scale analysis," in *ICCV*, 2009. 2

[3] X. Wang, X. Ma, and E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Trans. PAMI*, 2008. 2, 3, 8, 11

[4] J. Varadarajan and J. Odobez, "Topic models for scene analysis and abnormality detection," in *ICCV-12th IEEE International Workshop on Visual Surveillance*, 2009. 2

[5] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari, "What's going on? discovering spatio-temporal dependencies in dynamic scenes," in *CVPR*, 2010. 2, 3, 8, 11

[6] D. Blei and J. Lafferty, "Dynamic topic models," in *Proc. of the 23rd Int. Conference on Machine Learning*, 2006. 2

[7] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2006. 2

[8] T. Hospedales, S. Gong, and T. Xiang, "A markov clustering topic model for mining behavior in video," in *ICCV*, 2009. 2, 3, 8, 11

[9] J. Li, S. Gong, and T. Xiang, "Discovering multi-camera behaviour correlations for on-the-fly global activity prediction and anomaly detection," in *ICCV-12th IEEE International Workshop on Visual Surveillance*, 2009. 2, 11, 12, 13

[10] T. A. Faruquie, P. K. Kalra, and S. Banerjee, "Time based activity inference using latent dirichlet allocation," in *British Machine Vision Conference*, London, UK, 2009. 2

[11] J. Varadarajan, R. Emonet, and J. Odobez, "Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes." in *Procedings of the British Machine Vision Conference (BMVC)*, Aberystwyth, 2010. 2, 3, 8, 10

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, p. 1566–1581, 2006. 2

[13] X. Wang, K. Ma, G. Ng, and W. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric bayesian model," in *CVPR*, 2008. 2

[14] T. S. F. Haines and T. Xiang, "Delta-dual hierarchical dirichlet processes: A pragmatic abnormal behaviour detector." in *ICCV*, 2011. 2

[15] E. Jouneau and C. Carincotte, "Particle-based tracking model for automatic anomaly detection," in *ICIP*, 2011. 2

[16] O. Javed and M. Shah, "Tracking in multiple cameras with disjoint views," *Automated Multi-Camera Surveillance: Algorithms and Practice*, pp. 1–26, 2008. 2

[17] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *CVPR*, 2009, pp. 1988–1995. 2

[18] X. Wang, K. Tieu, and W. Grimson, "Correspondence-free multi-camera activity analysis and scene modeling," in *CVPR*, 2008. 3

[19] X. Wang, K. Tieu, and E. Grimson, "Correspondence-free activity analysis and scene modeling in multiple camera views," *IEEE Trans. on PAMI*, vol. 1, no. 1, pp. 893–908, 2009. 3

[20] P. Marmaroli, J.-M. Odobez, X. Falourd, and H. Lissek, "A Bimodal Sound Source Model for Vehicle Tracking in Traffic Monitoring," in *EUSIPCO*, 2011. 3, 14, 16

[21] C. Wang and D. Blei, "Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process," in *NIPS*, 2009. 8

[22] R. Emonet, J. Varadarajan, and J.-M. Odobez, "Multi-camera Open Space Human Activity Discovery for Anomaly Detection," in *AVSS*, Aug. 2011, p. 6. 9, 11

[23] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *CVPR*, 2011. 14, 18

[24] J. Varadarajan, R. Emonet, and J.-M. Odobez, "Bridging the past, present and future: Modeling scene activities from event relationships and global rules," in *CVPR*, Jun. 2012. 17

Rémi Emonet has a broad range of interests from software architectures to computer vision. He holds an engineer diploma in computer science from Ensimag (France). He did a Ph.D. thesis at INRIA and was a teaching assistant in computer science and software engineering. During his Ph.D., he worked on software architectures for pervasive computing and intelligent environments. In 2010, he joined the Idiap as a postdoctoral fellow, working on Bayesian models for unsupervised activity modeling from videos.

Jagannadan Varadarajan received the M.Sc. degree in Mathematics in 2003 and the M.Tech degree in computer science in 2005 from the Sri Sathya Sai University, India. He is currently pursuing the Ph.D. degree in the Idiap Research Institute, and Ecole Polytechnic Federal de Lausanne, Switzerland. He worked as Scientist at Hewlett-Packard Labs and GE Global Research, India, during 2005–2008. His research interests include Computer vision, Machine Learning.

Dr Jean-Marc Odobez (Ecole Nationale Supérieure de Télécommunications de Bretagne (ENSTBr), France, 1990; PhD at INRIA, Rennes University, France, 1994), was associate professor in computer science at the Université du Maine, Le Mans, France, from 1996 to 2001. He is now a senior researcher at both the IDIAP Research Institute and EPFL, Switzerland, where he directs the Perception and Activity Understanding team. His main areas of research are computer vision and machine learning techniques applied to multimedia content analysis, tracking, and human activity and behavior recognition. He is the author or coauthor of more than 100 papers in international journals and conferences. He is or was the principal investigator of 10 European and Swiss projects. He holds two patents on video motion analysis. He is the cofounder of the Swiss Klewel SA company active in the intelligent capture, indexing, and Web casting of multimedia conferences. He is a member of the IEEE, and associate editor of the Machine Vision and Application journal.