Published at the IEEE CVPR conference, Providence, 2012

Bridging the Past, Present and Future: Modeling Scene Activities From Event Relationships and Global Rules

Jagannadan Varadarajan[1,2], Rémi Emonet[1] and Jean-Marc Odobez[1,2]
[1] Idiap Research Institute, CH-1920, Martigny, Switzerland
[2] École Polytechnique Fédéral de Lausanne, CH-1015, Lausanne, Switzerland
{vjagann,remonet,odobez}@idiap.ch

Abstract

This paper addresses the discovery of activities and learns the underlying processes that govern their occurrences over time in complex surveillance scenes. To this end, we propose a novel topic model that accounts for the two main factors that affect these occurrences: (1) the existence of global scene states that regulate which of the activities can spontaneously occur; (2) local rules that link past activity occurrences to current ones with temporal lags. These complementary factors are mixed in the probabilistic generative process, thanks to the use of a binary random variable that selects for each activity occurrence which one of the above two factors is applicable. All model parameters are efficiently inferred using a collapsed Gibbs sampling inference scheme. Experiments on various datasets from the literature show that the model is able to capture temporal processes at multiple scales: the scene-level first order Markovian process, and causal relationships amongst activities that can be used to predict which activity can happen after another one, and after what delay, thus providing a rich interpretation of the scene's dynamical content.

1. Introduction

This paper deals with automatic scene analysis and behavior mining. As our primary application, we deal with data coming from videos taken from busy traffic scenes where several activities occur simultaneously with complex inter-dependencies. Our aim is to discover both local rules governing a sequence of activities and global rules controlling what is (dominantly) happening in the scene. For example, in many traffic scenes, one may find that activities are governed by global scene level rules (states) determined by the traffic lights. At a more local level, the activities are controlled by implicit rules such as the "right of way" that are followed, or by the sequence of trajectory segments that a pedestrian has to follow to reach a destination in a multi-camera set-up. Such analysis can have several applications for example in automatic stream selection and abnormality detection. Additionally, it can also be used as a prior for other tasks like tracking and pedestrian detection [13].

Addressing this problem is not straightforward due to various challenges. Firstly, at any given time, multiple activities are occurring in the scene, some happening independently of others, and some exhibiting complex temporal inter-dependencies. Secondly, usually, dependent activities do not merely depend on the immediate past as assumed in first order Markovian methods, but potentially on some activity farther in the past, causing random temporal lags between related activities.

Traditionally tracking based approaches were more popular for video surveillance and analytic tasks [9, 16]. While tracking isolates an object's behavior from the rest of the scene, it shows limited performance and incurs a high computational cost in crowded scenarios with multiple objects. Due to this, clustering based methods were used with simple low-level visual features to discover meaningful patterns of activities [18, 17].

An alternative approach is to use Topic models[1] like Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA). Applications of topics models in video scene analysis started as a niche area, but has received considerable attention recently due to its success in discovering semantically meaningful patterns from simple low-level visual features in an unsupervised fashion. Additionally, it brings in the powerful tools of probabilistic generative models enabling us to model complex real life phenomena.

Our work builds on top of the success of topic models applied to scene activity analysis. We propose a novel model called the **M**ixed **E**vent **R**elationship **M**odel **(MERM)**, that takes as input a binary activity matrix whose entries indicate the start of a fixed set of short-term temporal activities over time, and outputs both local and global scene level rules. The novelties of the model are the following:

[1] A class of generative models that deal with discrete data like word counts in text documents

1) the approach posits that activities can occur for different reasons: either as the start of an independent activity in a given context (scene state), or as the logical consequence of a previous activity occurrence (dependent case);

2) independent activity occurrence depends solely on the scene state, while each dependent activity can be associated with any activity in the past using transition tables and time lag probability distributions;

3) our scene-level state-space models several aspects of the scene: the scene state dynamics, the amount of activities that occur at any time instant and the proportion of which is independent (vs dependent), and what are the activities that most probably occur independently in a given state.

The model parameters are inferred using a collapsed Gibbs sampling technique. We evaluate our method extensively on several datasets from state-of-the-art papers. Both qualitative and quantitative results validate our model's effectiveness.

The rest of the paper is organized as follows. In the following section, we review relevant work done in the area of activity analysis. In Section 3, we present our model with the generative perspective and inference. Following this, experimental details (data, video preprocessing) are presented in Section 4. Our results along with analysis and conclusions are given in Sections 5 and 6 respectively.

2. Related work

Unsupervised activity analysis using topic models was first demonstrated in [14] where activity patterns were discovered from traffic scenes. In [12], activity based scene segmentation and abnormal event detection was done using PLSA [6]. These works used short clips as documents and quantized low level visual features (coming from optical flow, location, object size, etc.) as words, however the temporal information was ignored.

A slightly different approach in modeling scene activities was proposed in [11, 2], where temporal information was incorporated within each topic using explicit time variables to indicate the order of words within an activity, and the start of an activity within a document. The method captures relevant temporal patterns akin to trajectory segments, but they do not extract the intrinsic rules of the scene (e.g., due to signal cycles) that generate these patterns.

Few efforts were made to capture higher level semantics of the scene. Among them, [7] proposed a model that uses a Markov chain running between distinct global scene behaviors. The Markov-chain used correlates the global states and does not correlate the activities that form the global states. In [8], global rules of the scene were extracted by exploiting non-parametric methods like HDP-HMM [10] with an infinite mixture of HMMs with infinite number of states. However, as clearly stated in [8], the method only found a single HMM on all videos. Thus, in practice, it reduces to [7] with an automatic selection of number of HMM states. Furthermore, in [8], discovery of both local and global rules are handled separately. For local rule finding, they rely on an exhaustive exploration of activity combinations and on a comparison with predefined Markov templates which is both hard to compare to and not scalable.

Our approach differs from all of the above methods fundamentally. Our method posits that two types of activities can occur at any time instant: a) those that happen independently of the past (but with a higher probability depending on the scene context), and b) those that depend on previous event occurrences. These relationships are then jointly inferred in our model as global and local rules of the scene.

In our model, we use a dynamic (scene) state-space to capture the number of activities starting and the proportion of them occurring independently or dependently. While independent activities are triggered from the current state, dependent activities are decoupled from the state and can depend on any of the past occurring activities. Furthermore, the relationships among every pair of dependent activities are not only captured using transition probabilities, but also using distributions on time lag between their joint occurrences.

3. Model and Inference

In this section, we first introduce the model and then present the notations, the generative process and the inference method in more detail.

3.1. Model Overview and Notations

Fig. 1 shows an input matrix of observations representing activity occurrences. The main aspects of the proposed model are also illustrated in the figure.

The goal of our model is to capture which activity usually follows which one and with what delay. For example, we want to capture that activity 2 often occurs after activity 1 as in Fig. 1. This information is called a transition in the model and is noted as $\tau_{act1}(act2)$. Contrary to the widely used HMM-based methods, we want to capture precise time lag information in addition to the transition matrix. For example, we would like to model that activity 2 follows activity 1 with a time lag of 3 to 5 seconds. For each possible transition, a variable in the model describes the distribution of time lag noted as $\delta_{act1,act2}$.

Since, in the proposed MERM model an activity can be either dependent or independent of the past, we use filled-red boxes to represent independent activities and blue boxes for dependent activities in Fig. 1. Each dependent activity is supposed to be triggered by a single one in the past. This relation is represented with an arrow between the activities (from the past one to the dependent one).

More formally, each observation i has a variable s_i^t that is 1 when the activity is independent of any other. When an

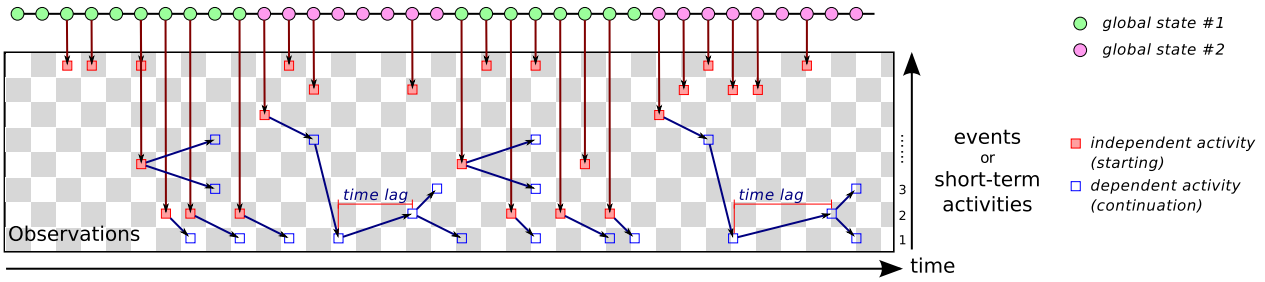

Figure 1. Schematic overview of the proposed model – Observations are temporally-localized occurrences of short term activities. Each observation occurs either independently or as a continuation of a previous observation. A scene-level state controls the amount of activities starting at a given instant, their type as well whether they are independent or not.

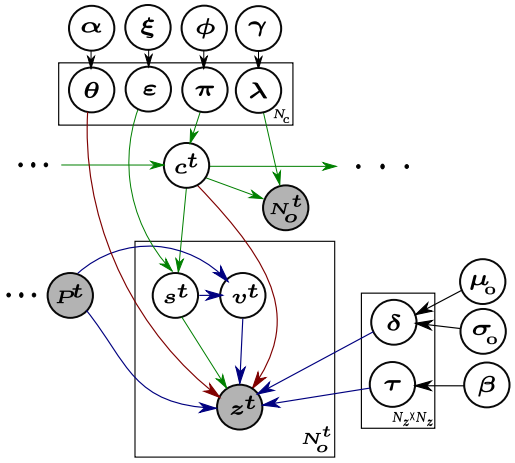

Figure 2. Graphical Model – green: links to generate all activities, blue: links to generate dependent activities, red: links to generate independent activities.

activity is dependent on a previous one, s_i^t is 0 and another variable v_i^t denotes the activity on which it depends.

A scene can be made of cycles with different phases where each phase produces different amounts and kind of activities or events. We model these phases by having a top-level HMM in which the current state controls 3 aspects of the model at the current time instant: the number of observed activities, the proportion of dependent and independent activities, and for the independent ones, the proportion of each kind of activities.

3.2. Generative Process

A video of a scene can be summarized using a fixed set of activities and their starting times. This can be represented as a binary activity occurrence matrix O containing T_d columns indicating time and N_z rows indicating the fixed set of activities. Every non-zero entry in the matrix O_i^t represents the start of an activity i at time t.

According to our model, at each instant t, the scene is in a state c^t among the N_c possible ones, depending on c^{t-1}, the state at time $t-1$. Each state $k \in [1..N_c]$ controls the number of activities that can occur using a Poisson distribution with parameter λ_k, the proportion of dependent or independent activities using a Bernoulli distribution with parameter ϵ_k and the set of independent activities using a Multinomial distribution with parameter θ_k. At each time t, the number of activities N_o^t is sampled from a Poisson(λ_{c^t}). Each of these occurrences O_i^t are associated with a binary decision variable $s_i^t \in \{0,1\}$ sampled from Bernoulli(ϵ_k) which decides if the activity is generated depending on one of the past occurrences or independently.

In cases where $s_i^t = 1$, we rely on the current state distribution to start an independent activity sampled from Discrete(θ_k). In cases where $s_i^t = 0$, the current observation depends on one of past activities at time $t' < t$. The set of past activities for time t is given by $P^t = \{O_i^{t'}\}_{i=1,t'<t}^{N_p^t}$. Practically we limit the dependency to a fixed temporal extent in the past $t - T_l \leq t' < t$. The notations used and the graphical model presentation of this model are provided in Table 1 and Figure 2.

The method of generating the activity occurrence matrix is described as follows.

- for each $k = 1, \ldots, N_c$ global states;
  - draw Poisson parameter $\lambda_k \sim \text{Gamma}(\gamma_1, \gamma_2)$
  - draw Bernoulli parameter $\epsilon_k \sim \text{Beta}(\xi_0, \xi_1)$
  - draw Multinomial parameter $\theta_k \sim \text{Dirichlet}(\alpha)$
  - draw state transitions $\pi_k \sim \text{Dirichlet}(\varphi)$
- for each activity type $z \in [1..N_z]$, draw transitions $\tau_z \sim \text{Dirichlet}(\beta)$;
- for each activity pair $z', z \in [1..N_z]^2$, draw lags $\delta_{z',z} \sim \mathcal{N}(\delta_{z',z} | \mu_0, \sigma_0^2)$
- for each t in $1, \ldots, T_d$
  - draw $c^t \sim \text{Discrete}(\pi_{c^{t-1}})$
  - draw a number $N_o^t \sim \text{Poisson}(\lambda_{c^t})$
  - for each i in $1, \ldots, N_o^t$
    - draw a binary value $s_i^t \sim \text{Bernoulli}(\epsilon_{c^t})$
    - draw $O_i^t \sim \text{Discrete}(\theta_{c^t})$ if $s_i^t = 1$,

Symbol	Description
α	Dirichlet prior on independent activities
β	Dirichlet prior on motif transitions
μ_0, σ_0^2	Hyper-parameters of Gaussian prior
$\gamma = \{\gamma_1, \gamma_2\}$	Hyper-parameters of Gamma prior
$\xi = \{\xi_0, \xi_1\}$	Hyper-parameters of Beta prior
φ	Dirichlet prior on state transitions
$\tau_z(z)$	Transition from activity z' to z
$\delta_{z',z}, \sigma^2$	Mean and a fixed variance on time lag between z' and z
θ_k	Distribution over activities set, for each state k
λ_k	Poisson parameter to select N_o^t, for each state k
ϵ_k	Bernoulli parameter to select s_i^t, for each state k
π_k	Global state transitions, for each state k
N_o^t	Number of activities at time t
P^t	Set of past activities for time t
N_p^t	Number of past activities for time t
T_k	Number of time instants explained by state k
T_l	Maximum lag for activity associations

Table 1. Notations used in the paper

- draw $v_i^t | P^t \sim \text{Uniform}(\frac{1}{N_p^t})$, if $s_i^t = 0$ and, draw $O_i^t \sim \mathcal{N}(t - t' | \delta_{z',z}, \sigma) \cdot \text{Discrete}(\tau_{z'})$, where $P^t(v_i^t) = z'$ and σ is a fixed variance.

Note that the fundamental difference with respect to existing approaches is the use of decision variables associated with each occurrence. This decouples or relates an activity to the current state of the scene. Furthermore, in the case where activities are related to past occurrences, their dependency is captured using a transition parameter and a distribution on the temporal lag between every pair of activities. This gives the flexibility in capturing activity relationships especially when their execution speeds have high variations.

In the generative process, we assume that when the decision variable is 0, there is at least one activity in the past that would take responsibility of generating the current activity. More precisely, when $s_i^t = 0$, we assume that $P^t \neq \emptyset$ and that there exists a past event z' such that $\tau_{z',z} \neq 0, \delta_{z',z} \neq 0$. When $P^t = \emptyset$, we expect that such a state of the scene is captured in the global state variable hence generating only $s_i^t = 1$. Please see the inference method on how this issue is addressed by jointly sampling s_i^t, v_i^t.

A similar use of binary decision variables can be see in text mining domain, where it is used to decide if words need to be associated to form phrases [4, 15] or not. However, unlike in text, where there is a single observation at anytime, videos have multiple event occurrences at any instant with large temporal variations making this association a complex problem. It is also interesting to note that several types of Hidden Markov Model (HMM) can be derived from our MERM model. For instance, when the number of activities N_o^t is 1, it reduces to a kind of HMM where observations at each time instant can be either dependent or independent of the current state. Similarly, when s_i is always set to 1, it reduces to another kind of HMM with the states generating observations from a Discrete distribution.

3.3. Model Inference

As is the case with many hierarchical Bayesian models [5], exact inference for our model is intractable. But, thanks to conjugate pairs like Poisson-Gamma, Dirichlet-Multinomial and Beta-Bernoulli and Normal-Normal, it is possible to derive a collapsed Gibbs sampling algorithm by integrating out the parameters $\{\pi, \lambda, \epsilon, \theta, \tau, \delta\}$. The algorithm proceeds by iteratively sampling the decision variable s_i^t, indicator variable v_i^t for each observation O_i^t conditioned on all other variables, parameters and hyper-parameters. The state indicators c^t are sampled for each time instant conditioned on rest of the variables.

Since each occurrence also gives its occurrence time implicitly, we will drop the t associated with s_i^t and denote it by s_i except in places where time needs to be mentioned explicitly. We will also use O, S, V to refer to the set of occurrences, their corresponding selector variables, and indicator variables. $O_{\neg i}, S_{\neg i}, V_{\neg i}$, will indicate all the occurrences, selector variables and the indicator variables except the i^{th} one and hp refers to the set of hyper-parameters set: $\{\varphi, \gamma, \xi, \alpha, \beta, \mu_0, \sigma_0^2\}$. (For complete details about derivation of these equations, please see the additional material).

For a given observation i, we re-sample s_i and v_i jointly. For the case where $s_i = 1$ (independent event), v_i is not meaningful and we obtain the following sampling probability:

$$p(s_i = 1, v_i | S_{\neg i}, V_{\neg i}, O, C, hp) \propto \frac{q_{-i,k}^{(z)} + \alpha}{q_{-i,k}^{(.)} + N_z \alpha} \cdot \frac{l_{-i,k}^{(1)} + \xi_1}{l_{-i,k}^{(.)} + \xi_0 + \xi_1} \quad (1)$$

For the case of an observation that depends (i.e. $s_i = 0$) on the j^{th} one ($v_i = j$) that is of type z', we have the following sampling probability:

$$p(s_i = 0, v_i = j | O, S_{\neg i}, V_{\neg i}, c_i = k, C_{\neg i}, hp) \propto \frac{1}{N_p^t} \cdot \frac{l_{-i,k}^{(0)} + \xi_0}{l_{-i,k}^{(.)} + \xi_0 + \xi_1} \frac{r_{-i,z'}^{(z)} + \beta}{r_{-i,z'}^{(.)} + N_z \cdot \beta} \mathcal{N}_{\delta_{z',z},\sigma}\left(f(P^t(j), O_i^t)\right) \quad (2)$$

In the above expressions $q_{-i,k}^z$ is the count of number of times motif z is observed with state k when $s_i^t = 1$ removing the current observation. Similarly, $l_{-i,k}^1$ and $l_{-i,k}^0$ are

the counts of $s_i = 1$ and $s_i = 0$ appearing with state k barring the current observation. $r^{(z)}_{-i,z'}$ is the count of event z appearing after z' and $f(P^t(j), O^t_i)$ is the temporal delay between $P^t(j) = z'$ and $O^t_i = z$. The mean for each pair of events $\delta_{z',z}$ is the estimated posterior mean obtained from all time lags between associations of z' and z except the current one and σ is a fixed variance in the lag.

For re-sampling the global state c^t at time t, the sampling depends on three main factors coming from the links in the graphical model. These factors are relatively complicated and provided in the additional material.

4. Experiments

Datasets – We evaluate our method on 4 different video datasets of busy traffic scenes used in state of the art papers. The datasets are broadly of two kinds: a) controlled by traffic lights and b) uncontrolled. The MIT data [14] (1.2 hrs) and QMUL Junction dataset [7] (1 hr) are footages of two different busy 4 road junctions controlled by traffic lights. They have a variety of activities such as vehicles moving in different directions and pedestrians crossing the road. For the uncontrolled case, the Far-field dataset [11] (2 hr) was used. This video is over a three road junction with mainly vehicular activities but with significant temporal variations. The ETH dataset [8] (in additional material) is 50 minutes long and has both structured and unstructured activities.

Video to Events – As mentioned earlier, our method takes a sparse binary activity occurrence matrix as input where the non-zero entries correspond to activity starts. In this work we use the Probabilistic Latent Sequential Motifs (PLSM) method proposed in [11] that takes a temporal document as input and produces sequential activity patterns called motifs as output.

More precisely, as a first step background subtraction is performed on the video and the KLT tracker is used to obtain optical-flow features at every point in the foreground. The optical flow features are quantized using nine different bins i.e., 8 bins for the 8 cardinal directions and one bin for static pixels. The location is quantized into small grids of 2×2 non-overlapping pixels. PLSM, first uses a dimensionality reduction step to obtain spatially localized activity patterns whose amount of presence at each time step is used to create a temporal document. By applying PLSM on these temporal documents we obtain dominant activities of the scene. The advantage of using PLSM is that it gives the starting times of the discovered dominant activities which are directly used in creating our binary occurrence matrix.

Settings – Based on the complexity of the scene and variety of activities occurring in each of the datasets we extracted 25, 20, 20 and 15 sequential patterns from the MIT, QMUL Junction, ETH and Far-field datasets, all with a fixed maximal duration of 5 or 10 seconds. Note that this places only an upper limit on the duration of the patterns that we obtain. Each column of the activity matrix indicates a 1 second temporal resolution. As we have no a-priori information on any of the parameters, we used non-informative symmetric Dirichlet distributions for $\{\alpha, \varphi, \beta\}$. Similarly, we set equal values to the Beta hyper-parameters $\xi_0 = \xi_1 = 1$ and the same follows for the Gamma distribution hyper-parameters too. Note that our events have a maximal duration of 5 seconds. In order to accommodate large temporal variations between dependent activities we used a flat Gaussian distribution with $\mu_0 = 5$ and $\sigma_0^2 = 4$. The maximum lag T_l is set to twice the duration of activities i.e, 10 or 20. We used 90% of the duration of each video mentioned for training our MERM model. Gibbs sampling was run for a sufficiently long number of iterations (1000 for each of them). However, in all our experiments we found that the sampler reached stationarity after 600 iterations.

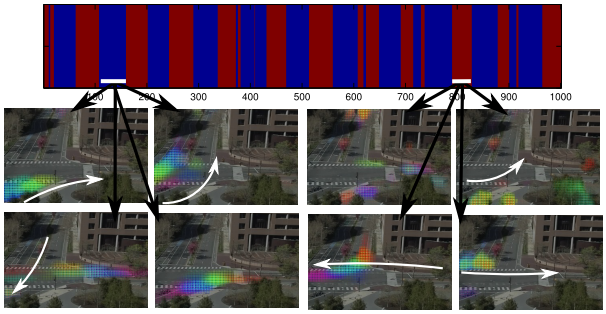

Figure 3. Two scene states obtained from MIT data shown for 1000 seconds. The two states correspond to distinct periods of the traffic signals cycle. The time indicated in blue correspond to motion in the north-south direction. The red region correspond to east-west movement.

5. Results

5.1. Global rules

In our model the global rules are characterized by the scene-level states, their transition probabilities, and their attributes. These correspond to the parameters $\{\pi, \theta, \epsilon, \lambda\}$. Here, we show global rules obtained from MIT and QMUL Junction datasets.

MIT dataset – On the MIT dataset, we used 2 states in our model to produce the results shown in Figure 3. We also experimented with 3 and 4 states and obtained comparable results with a finer segmentation of the scene activities.

The two states inferred by our model correspond to the two main periods in the traffic cycle. A first state (in blue) corresponds to vertical movements and stopped cars on the horizontal lanes. Similarly, the second state (in red) captures horizontal car movement while cars on vertical lanes are waiting (only one of these activities is shown due to space restrictions). We see that the scene periodically fol-

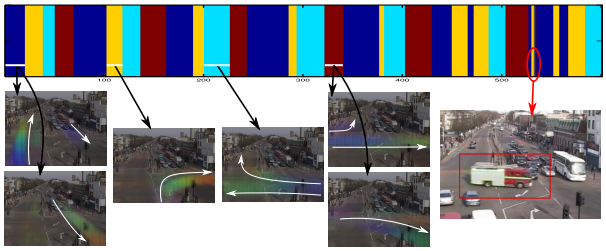

Figure 4. Four scene states obtained by our model from the QMUL Junction dataset. The four states represented by the colors blue, cyan, yellow and red are plotted for 600 seconds of video. Lower part: dominant activities found to occur at each of the 4 states. Red ellipse: an abnormality where an emergency vehicle interrupts the traffic momentarily.

lows the alternating traffic signals, except for some rare glitches.

QMUL Junction dataset – With a-priori knowledge about the scene activities, we applied MERM model to extract 4 global states from QMUL Junction dataset. The colored pattern (blue, cyan, yellow and red) in Fig. 4 top, shows the discovered states of the scene over 10 minutes of video. Below the scene states we see the top ranking independent activities of the four states. These activities correspond well to vehicles moving in the vertical directions (blue), moving from bottom and turning towards the right (yellow), from left to right (red) and right to left (cyan) respectively.

The repetitive pattern of colors (states) clearly shows that our model properly captures the periodic cycles of the scene that are due to traffic lights. A notable exception in the periodic repetition of the states is circled in the right part of Fig. 4. This is an unusual event when an fire engine crosses from left to right, freezing all other traffic.

5.2. Local rules

Our model captures local rules in the form of transitions from one activity to another, with a time lag and a fixed-variance normal distribution. We can represent the set of local rules in the form of a graph with (blue) edges annotated with a weight and a mean lag information as done for example in Fig. 5 (detailed later). On such a figure, we can also show the proportion of independent occurrences for each activity as some sourceless (red) edges. As independent activities can be caused by different scene-level states, the sourceless edges are annotated with the scene-level state (c_i) that cause them. For space and readability reasons we filtered out low frequency edges in the illustrations.

Far-field dataset – Figure 5 shows the graph of activities recovered by our model for the Far-field dataset with PLSM motifs of 5 seconds duration. In this dataset, 13 of the 15 activities have been found to have notable dependencies. Interestingly, Fig. 5 properly exhibit two independent subparts in the graph. The two subgraphs correspond to the two main vehicle directions.

The upper part of Fig. 5 shows a chain a-b-c-d-e for a vehicle coming from the right and disappearing on the top of the image. Only activities a and b have been found to be spontaneous starts of activities. The sum of the lags δ of each subgraph also gives the duration of the full trajectory of a vehicle in the scene which is around 15 to 20 seconds. To account for some vehicles slowing down or stopping at this location due to a single-lane tunnel, we see that d and e activities have self loops.

To account for possible variations in vehicle speed, some "shortcut" transitions are also captured. For example, we see that it is common to do directly b-d in 5.4 seconds instead of doing b-c-d in 6.1 seconds.

The lower part of Fig. 5 captures the other direction for cars. Starts are spread on the three activities f, g, and h. The loop on state f (same for g and h) is explained by the scene: cars coming from the top tend to group together as there is a single-lane tunnel where cars cannot cross and have to wait for the other direction to finish passing. A similar graph with motifs of 10 seconds duration (used for experiments in sec 5.3) is provided in the additional material.

MIT dataset – On the MIT dataset, 24 out of 25 activities have been found to have dependencies. Activities with high dependencies are shown in Fig. 6. Among them, the six activities on the right corner of the figure are only self dependent and mostly correspond to stopped cars. Dependent activities j-k also correspond to tree motion.

Three activities (lower-right corner) are found to be starting and self-looping: they all correspond to examples of cars starting. A car starting is an activity that repeats itself because multiple cars are usually waiting for the green light and start successively.

In addition to repeating static activities, the model captures trajectories made of multiple activities such as h-i and d-e-c. Some interesting soft rules are also captured like in f-g: cars coming from the left (g), often turn just after a car coming in the opposite direction has passed.

5.3. Evaluation

To objectively measure the performance of our model we used an activity prediction task as proposed in [2]. Here, we predict future observations using parameters inferred from only partial observations. This is akin to computing predictive perplexity that is popularly used in language models [1]. More precisely, in [2], at each time t a window of past activities $W_{\{t-T_m \leq t' < t\}}$ are used to predict the observations at time t, $p^{pred}(W_t)$. This is then compared to the ground truth observations $p^o(W_t)$ using the normalized likelihood

$$PLL = \frac{\sum_w p^o(W_t) \log(p^{pred}(W_t | W_{t-T_m \leq t' < t}))}{\sum_w p^o(W_t)} \quad (3)$$

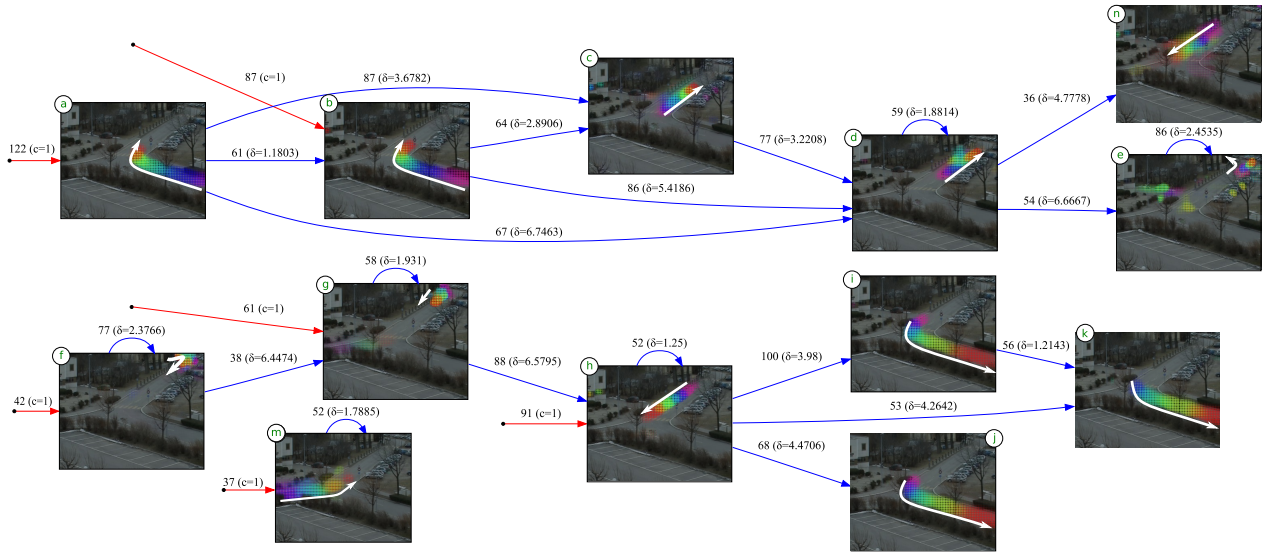

Figure 5. Event relationships from Far-field data. Low frequency edges are not shown (those with counts below 30). Blue edges: transitions with weight and lag δ. Red edges: independent activity with weight and causing state (here with only 1 state). 80% of the occurrences were dependent activities.

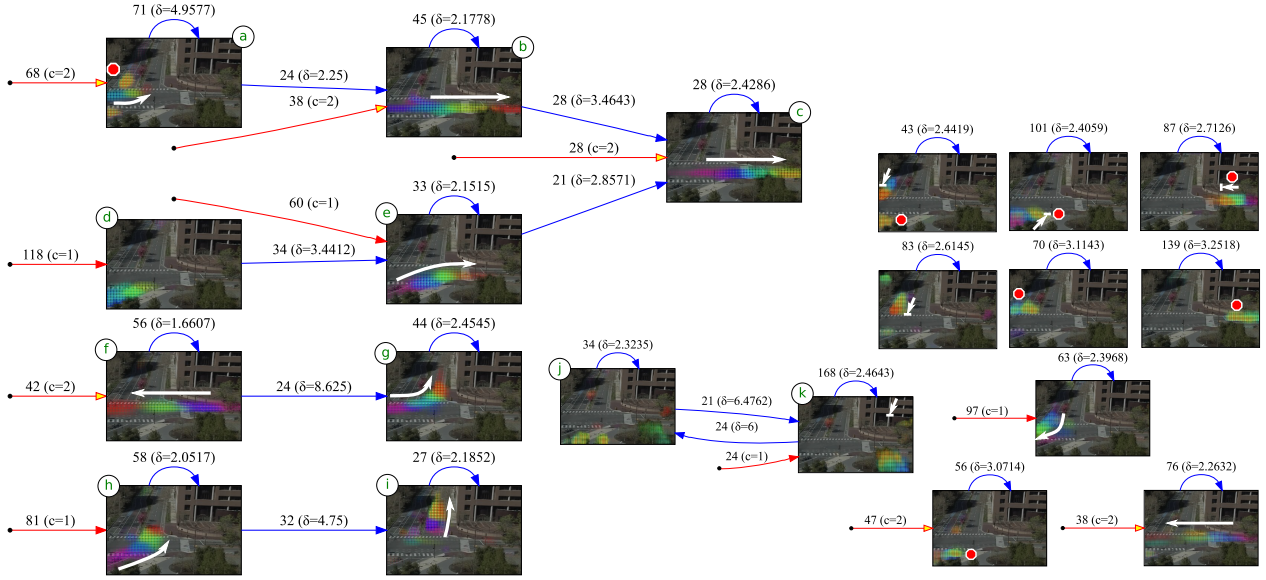

Figure 6. Event relationships from MIT dataset. Low frequency edges with counts below 20 are filtered. Blue edges: transitions with weight and lag δ. Red edges: independent activity with weight and causing state. 75% of the occurrences were dependent activities.

In [2], observations from a subwindow of $T_m = 29$ duration were used to predict observations at time $t = 30$. Here, the predictions are strictly based on past observations and not based on any global rules of the scene. Since MERM learns the global rules and inter-activity dependencies, we supplement predictions from [2] with MERM as prior. So, for each subwindow of 29 instants MERM is applied to predict the possible activities for time $t = 30$ using its properties (number of events, proportion of dependent events) and the past observations (transition and lag).

Experiments were carried out on the Far-field data. Fig. 7 shows average predictive log-likelihood (PLL) computed from these models using a 10 fold cross validation procedure. The plots in 7 show the prediction performance for first and second future time steps $(t, t + 1)$ obtained from our MERM model with one state, simple PLSM [2] and from [7] (Topic-HMM). X-axis indicates the number of PLSM activities with 10 second duration or the number of behavioral states in Topic-HMM. We observe that using MERM over PLSM outperforms both simple PLSM

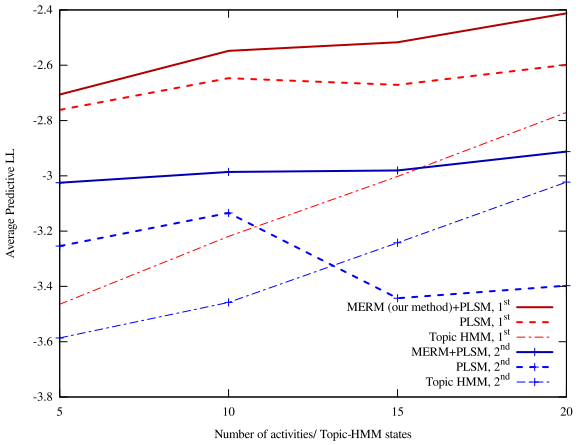

Figure 7. Prediction accuracy – Average predictive log-likelihood, PLL for the first or the second future time steps.

and Topic-HMM for both the time steps. The impact of MERM over PLSM, due to learning scene level rules is significant in case of the second time step, where it performs poorly. We also observe that Topic HMM performs less than PLSM initially, and improves when the number of states is increased.

6. Conclusion

In this work, we presented a novel model (MERM) that accepts a binary activity matrix and discovers temporal relationships between activity pairs as well as global rules dictating the scene. The main novelty comes from our observation that activities need not be strictly dependent on the current state of the system but also on any activity in the past. We proposed an efficient Gibbs sampling algorithm to infer the latent variables and the parameters. The hard task of attributing an activity to the past or to the global scene state is solved thanks to the joint sampling of the latent variables. We evaluated our method on a variety of scenes containing complex activity patterns. The results demonstrate our model's ability to capture both local rules (event-event relationships) and global rules (event-state relationships). Future plans include applying our model on data from multiple-cameras as in [3]. This could be useful in identifying independent activities and infer inter-camera activity relationships along with their temporal lags.

Acknowledgements

The authors gratefully acknowledge the financial support from the Swiss National Science Foundation (Project: FNS-198,HAI) `www.snf.ch/E` and from the 7th framework program of the European Union project VANAHEIM (248907) `www.vanaheim-project.eu` under which this work was done.

References

[1] D. M. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, pages 993–1022, 2003.

[2] R. Emonet, J. Varadarajan, and J.-M. Odobez. Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model. In *CVPR*, 2011.

[3] R. Emonet, J. Varadarajan, and J.-M. Odobez. Multi-camera open space human activity discovery for anomaly detection. In *AVSS*, 2011.

[4] T. Griffiths, M. Steyvers, and J. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211–244, 2007.

[5] G. Heinrich. Parameter estimation for text analysis. Technical report, 2004.

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

[7] T. Hospedales, S. Gong, and T. Xiang. A Markov clustering topic model for mining behavior in video. In *ICCV*, 2009.

[8] 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.

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

[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[11] J. Varadarajan, R. Emonet, and J. Odobez. Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. In *BMVC*, 2010.

[12] J. Varadarajan and J. Odobez. Topic models for scene analysis and abnormality detection. In *IEEE IWVS*, 2009.

[13] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *CVPR*, 2011.

[14] X. Wang, X. Ma, and E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):539–555, 2009.

[15] X. Wang, A. Mccallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *IEEE Int. Conference on Data Mining*, 2007.

[16] X. Wang, K. Tieu, and E. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE T-PAMI*, 1(1):893–908, 2009.

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

[18] G. Zen and E. Ricci. Earth mover's prototypes: a convex learning approach for discovering activity patterns in dynamic scenes. In *CVPR*, 2011.