Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Discovering Spatio-Temporal Relationships Among Activities in Videos Using a Relational Topic-Transition Model

Dalwinderjeet Kular and **Eraldo Ribeiro**
Florida Institute of Technology
Melbourne, U.S.A.

Abstract

Discovering motion activities in videos is a key problem in computer vision, with applications in scene analysis, video categorization, and video indexing. In this paper, we propose a method that uses probabilistic topic modeling for discovering patterns of motion that occur in a given activity. Our method also identifies how the discovered patterns of motion relate to one another in space and time. The topic-modeling approach used by our method is the relational topic model. Our experiments show that our method is able to discover relevant spatio-temporal motion patterns in videos.

1 Introduction

Video recordings are a common part of modern life. Daily, people record videos using mobile devices, surveillance cameras record activities in public places, YouTube™ users upload 300 hours of video every minute[1]. These recordings contain a great deal of information. However, extracting information from videos is a challenging task that for many applications require people to watch videos for hours while annotating the necessary information. Clearly, an automated solution to video analysis is desirable.

This paper addresses the problem of automatically extracting information from videos. Here, we focus on the analysis of activities that are formed by combining basic motions. Examples of such activities include traffic scenes, rugby players passing a ball, and people moving in a hotel lobby (Figure 1). Analyzing these activities is usually simple for humans to perform. In many cases, we can quickly discover the basic patterns of motion that forms an activity as well as the temporal sequence in which those patterns occur. For activities with complex motion dynamics, as we watch them for a sufficient amount of time, we eventually notice the emergence of basic motion patterns.

Most methods for analyzing motion activities use a two-step solution. The first step discovers the basic motion patterns that compose the activity. The second step links the discovered motion patterns into sequences to create a model of the activity. While these two steps are conceptually simple to define, their implementation is challenged when only local motion measurements are available such as optical flow or trajectory fragments. While being the simplest motion cues that can be measured from video, optical flows are essentially instantaneous velocities of moving pixels. This type of measurement is noisy, sparse, and most importantly, does not provide long-duration (i.e., contiguous) motion information or even a direct connection to a specific object (i.e., tracking). Given the inherent uncertainty of motion measurements such as optical flow and trajectory fragments, a number of statistical-modeling methods borrowed from the field of document analysis have been successfully applied to the problem of analyzing activities from video data. Examples of these techniques include the latent semantic analysis and latent Dirichlet allocation. For example, Wang, Ma, and Grimson (2009) grouped motion events into activities using hierarchical Bayesian models. However, these methods extracted spatially co-occurring motion events and neglected temporal relationships. Hospedales, Gong, and Xiang (2009) developed a Markov Clustering Topic Model based on the Latent Dirichlet Allocation (Blei, Ng, and Jor-

Figure 1: Activities formed by combinations or sequences of various basic motions. (a) A traffic scene with traffic moving downwards and upwards. (b) A player passing of the ball to another player. (c) People moving in a hotel lobby.

[1] http://www.statisticbrain.com/youtube-statistics

Figure 2: Traffic-intersection scene. Scene consists of multiple actions, motion patterns, and activities. Traffic patterns include vertical traffic from both directions, vertical turns, horizontal traffic flow, and back to vertical traffic flow.

dan 2003). Their model clustered spatial visual events into activities and identified the temporal dynamics of the visual events. These methods require the number of motions to be learned as input, which may lead to some motions being missed. In Hospedales, Gong, and Xiang (2009), the use of a single Markov chain for learning occurrence cycles may result in incomplete activities.

In this paper, we propose a *Relational Topic-Transition Model* for discovering the main basic motions that composed the activity as well as detecting how these discovered motions relate to one another in both time and space. For example, given a video of a traffic scene (Figure 2), we want to determine: the driving lanes, the sequences of motions happening in these lanes (e.g., moving forward, moving forward then right turn, turning about). In addition to finding major flow directions, we also want to discover the temporal synchronization among the motions (i.e., traffic flows that move sequentially and in co-occurrence). Finally, we want to do all that by taking as input just local velocity measurements of the scene's moving objects (i.e., optical flow).

Our method is based on the concept of motion patterns (Saleemi, Shafique, and Shah 2009). We consider a *motion pattern* to be a spatially and temporally coherent cluster of moving pixels. These clusters are connected together into a relatively long spatio-temporal region describing the motion of objects from one place to another in the scene. The dashed curves in Figure 2 show some of the motion patterns for a traffic-intersection activity. Our method combines a Relational Topic Model and an Activity Transition Model. The Relational Topic Model is a text-based document-analysis model proposed by Chang and Blei (2009), which we adapted to analyze videos. The Relational Topic Model discovers actions and motion patterns in a single step while the activity transition model identifies activities (i.e., the temporal synchronization among motion patterns).

We tested our method on four publicly available traffic-flow datasets (Hospedales, Gong, and Xiang 2012). To our knowledge this is the first time Relational Topic Modeling (Chang and Blei 2009) is being used for analyzing videos.

2 Relational Topic Model for Videos

We adapted the Relational Topic Model (RTM), a text-based document analysis model, for analyzing video scenes. RTM is a hierarchical generative model proposed by (Chang and Blei 2009). It generates hidden themes from a set of documents for generating new documents, explaining existing documents, and the relationship shared among those documents. These hidden themes are known as *topics*. Topic is a distribution over a fixed-vocabulary of words. In RTM, for generating a document, first a topic proportion is sampled from a Dirichlet distribution. For each word in the document, a topic is sampled from the topic proportion, and then a word is drawn from the sampled topic distribution. These steps are the same as the used in the Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003). In addition to generating the documents, RTM generates a *binary link probability* among the documents. This binary link probability identifies whether a pair of documents are connected. The connection between the documents means that the documents are generated using similar topics. In our approach, we used LDA, part of RTM, to learn actions and action dependency over the links to learn multiple activities. In Figure 3, we show how RTM connects clips of the video.

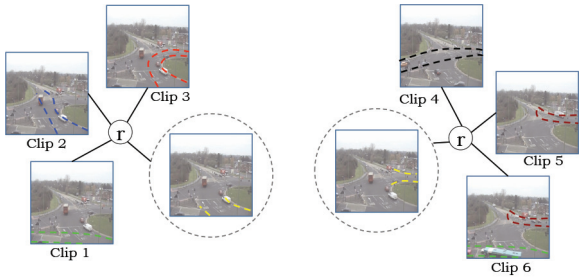


Figure 3: A traffic scene observed at a roundabout. Each clip is represented as a set of actions and clips are linked to each other on the bases of similar actions.

In our approach the documents and topics have a different interpretation than that used for document analysis. Here, the *topics* were the basic parts of the motion flow (e.g., basic connected regions of pixels displaying a coherent motion in some direction). We call these topics *actions*. The *documents* were the short video clips (i.e., a set of a few consecutive frames) of fixed duration. Finally, the *words* in each document were the basic visual-events that occurred with certain frequency depending on the actions that form the video clips. The basic visual-events (i.e., words) were part of a fixed-length vocabulary.

We call a *motion event* as any sufficiently fast motion happening in the scene at a given time instant. These motions are the instantaneous velocities of objects moving in the scene. The detection of motion events is done as follows. First, we divide the video into a set of non-overlapping clips, $\mathcal{C} = \{c_1, \ldots, c_M\}$, where each clip has a fixed number of image frames, i.e., $c_i = \{I_1, \ldots, I_N\}$. Then, we compute the optical flow between pairs of all consecutive frames within the clip. This calculation produces a set of 2-D vector fields of motion events $\mathbf{v}(\mathbf{x}) = (u, v)^\mathsf{T}$ for each clip, where u and v are the components of the velocity vector at pixel $\mathbf{x}$.

To create a suitable statistical analogy from a video clip to a document in RTM, we summarize the clip's motion information as a two-dimensional matrix E whose elements contain the frequency of all sufficiently fast motion events

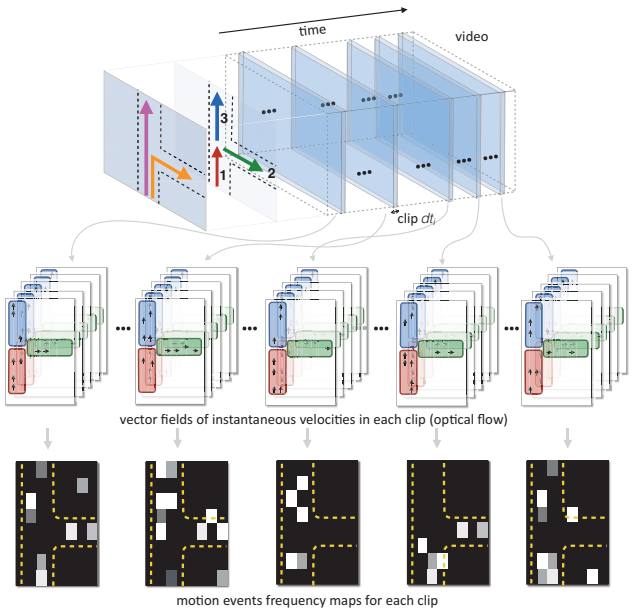


Figure 4: **Feature extraction**. An activity formed by two motion patterns: a forward motion (pink) and a right-turn motion (orange). Motion patterns are joined short-duration motion components (red, green, and blue arrows). We divide the video into consecutive clips of duration dt_i. For each clip, we calculate the optical flow between consecutive pairs of frames to generate a set of velocity vector fields of moving objects. Pixel locations with high-magnitude motion are marked in each vector field. The frequency of these pixels is recorded into a motion-event map that is created for each clip. Both the velocity fields and the motion-event maps are the input to our method.

occurring at each pixel location during the clip. For a clip c of N image frames, we have:

$$E_c(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(\|\mathbf{v}(\mathbf{x})\| > \tau\right), \tag{1}$$

where the indicator function $\mathbb{1}(\cdot)$ returns 1 when its argument is true and 0 otherwise, and τ is a pre-defined threshold value. By considering only motions above certain magnitude, we filter out noise and less-relevant motions. Figure 4 (bottom row) illustrates the motion-frequency maps summarizing the motion in each video clip. These maps are the input to our RTM model for discovering actions (i.e., motion-pattern parts).

We assume that a motion activity is composed by sequences of motion patterns as well as sets of co-occurring motion patterns. Motion patterns have been used before for activity analysis (Saleemi, Shafique, and Shah 2009). Here, we consider motion patterns as spatio-temporal clusters of moving pixels describing the flow of objects along a preferential path. To form motion patterns, our method clusters optical-flow data, without information about individual objects. As a result, different motion patterns that share a spatial path in the scene will be necessarily split into parts. These parts, when temporally connected together along a spatial path, form a motion pattern. The motion patterns in Figure 4 are composed by three simpler motion parts shown as arrows in red, green, and blue (also labeled by numbers 1, 2, and 3). The motion pattern corresponding to the longer vertical forward motion (i.e., pink curve) is composed of the two shorter vertical forward motions shown by arrows 1 (red) and 3 (blue). The motion pattern corresponding to the vertical right turn is composed of a short vertical-forward motion (i.e., action 1 in red) followed by a horizontal forward motion (i.e., action 2 in green). We call these motion parts *actions*.

The RTM's generative process is as follows:

1. For each clip c,
   (a) Action proportions are drawn from Dirichlet distribution, $(\theta_c|\alpha) \sim Dir(\alpha)$.
   (b) For each visual-event, $e_{n,c}$,
      i. Action $a_{n,c}$ is drawn from the action proportion, $a_{c,n}|\theta_c \sim Multi(\theta_c)$.
      ii. Visual-event $e_{n,c}$in clip c is then sampled from the corresponding action sampled from action-proportion distribution $(\beta_k, a_{n,c}, e_{n,c}) \sim p(e_{c,n}|a_{n,c}, \beta_{1:K})$.
2. For each pair of clips $\{c, c'\}$ the binary link indicator is generated,
   (a) Binary indicators are distributed according to the actions used for clips generation, $(\eta, a_{c,n}, a_{c',n}, r_{c,c'}) \sim p(r_{c,c'}|a_{c,n}, a_{c',n}, \eta)$, where $p(r_{c,c'}|a_{c,n}, a_{c',n}, \eta) \sim \psi(.|a_{c,n}a_{c',n}, \eta)$ and ψ provides binary probabilities.

The complete generative model for RTM is:

$$p(\beta_k|\phi) \sim Dir(\phi), \tag{2}$$
$$p(\theta_c|\alpha) \sim Dir(\alpha), \tag{3}$$
$$p(a_{c,n}|\theta_c) \sim Multi(\theta_c), \tag{4}$$
$$p(e_{c,n}|a_{c,n}, \beta_{1:K}) \sim Multi(\beta_{a_{c,n}}), \tag{5}$$
$$p(r_{c,c'}|a_{c,n}, a_{c',n}, \eta) \sim \psi(.|a_{c,n}, a_{c',n}, \eta). \tag{6}$$

The full joint distribution of variables $\{e_{c,n}, a_{c,n}, r_{c,c'}\}_1^M$ and parameters β, θ given the hyper-parameters ϕ, α, η is:

$$p(\{e_{c,n}, a_{c,n}, r_{c,c'}\}_1^M, \beta, \theta, |\phi, \alpha, \eta) = \prod_{cn} \underbrace{p(e_{c,n}|\beta, a_{c,n})}_{\text{visual-event}} \prod_{cn} \underbrace{p(a_{c,n}|\theta_c)}_{\text{action}} \prod_{c,c'} \psi \underbrace{p(r_{c,c'}|a_{c,n}, a_{c',n}, \eta)}_{\text{relationship}} \times \prod_k \underbrace{p(\beta_k|\phi)}_{\text{actions}} \prod_c \underbrace{p(\theta_c|\alpha)}_{\text{action proportion}}. \tag{7}$$

The function ψ was distributed over the relationship shared between two clips. This distribution function was conditioned on the actions $\{a_{c,n}, a_{c',n}\}$ that generated clips $\{c, c'\}$.

For inference and parameter estimation, we used Gibbs Sampling algorithm. Gibbs Sampling algorithm belongs in the Markov Chain Monte Carlo (MCMC) framework. In RTM we estimated the action proportion per clip θ_d, the action-event distribution β_k, action index assignments for

each visual-event a_i, and action dependency over links. Gibbs Sampling can derive conditional distribution for every latent variable. Additionally, if an action a is known, then we can calculate action proportion and action-event distribution from action index assignment, and from action proportion of clips we can estimate an action dependency matrix. Therefore, inference begins with the assignments of visual-events to actions by integrating out parameters β, θ ((Griffiths and Steyvers 2004)). The Gibbs sampler for RTM computes the probability of an action a being assigned to a visual-event e_i, given the rest of the action assignments to visual-events:

$$p(a_i|\mathbf{a}_{-i}, \mathbf{e}), \tag{8}$$

where $\mathbf{a}_{-i}$ means all the action assignments except for a_i. For establishing dependency between clips based on actions we have:

$$\psi_N(r=1) = \exp(-\eta^T(\overline{a}_{c,n} - \overline{a}_{c',n}) \circ (\overline{a}_{c,n} - \overline{a}_{c',n}) - \nu), \tag{9}$$

where $\overline{a}_{c,n} = \frac{1}{N_c}\sum_n a_{c,n}$ and the notation $\circ$ denotes the Hadamard (element-wise) product. The link probability depends on a weighted squared Euclidean difference between the two action distributions.

2.1 Activity Transition Model

The Activity Transition Model discovers the sequence in which the motion patterns are occurring in the scene. Our approach was inspired by (Stauffer 2003)'s Tracking Correspondence Model for stitching tracklets together to form tracks if they belong to the same object and estimating sink and source of the tracks. In our method, motion patterns were treated as tracklets and activities were determined from a pairwise motion pattern's transition's likelihood matrix. Given s motion patterns, we estimated the transition matrix, Γ, where γ_{ij} is 1 if motion pattern, l_j was the first correspondence instance to occur after motion pattern l_i in the video scene. Then we defined a pairwise transition likelihood matrix, T, such that $T_{ij} = p(\gamma_{i,j}|L)$. The pairwise activity transition likelihood matrix is $s \times s$ matrix. This matrix had all the activities and contains pairwise transition probabilities.

3 Experiments

3.1 Synthetic Dataset

To show our method's capabilities and compare it with existing models, we generated a synthetic dataset. Synthetic dataset facilitates with a controlled situation where to interpret correct results is easier. Here, the synthetic dataset represents a very similar scenario to real-life traffic observed at a four-way traffic light intersection. In our synthetic dataset we have traffic flowing in horizontal and vertical directions and traffic taking turns (horizontal right and vertical left turn). The order in which traffic occurs is: horizontal crossing, horizontal right turn, vertical crossing, vertical left turn, and back to horizontal traffic flow (as shown in Figure 5). As we know four activities are occurring in the scene, we set our $K = 4$ (Hospedales, Gong, and Xiang 2009) and hyperparameters as $\{\alpha = 0.25, \beta = 0.05, \eta = 0.5\}$.

We have generated $1,000$ ordered clips from a generative model (a sampled dataset is shown in Figure 5), with each clip flattened into a 5×5 matrix. As shown in Figure 5, the bright square represents the presence of the traffic flow, whereas the black square means the absence of the traffic flow. Also, brightness and dullness of the cells mean the frequency of traffic being observed in that cell and these cells were considered visual-events. Therefore, the size of the vocabulary for synthetic data is 25. In this dataset, four lanes are formed i.e., traffic observed from the left, right, top, and bottom directions (as shown in Figure 6.a). The combination of these lanes generates horizontal traffic, vertical traffic and turns (as shown in Figure 6.b). Overtime, these crossing are occurring in the following pattern, i.e., horizontal traffic, horizontal turn, vertical traffic, and vertical turn (as shown in Figure 6.c). The actions, motion patterns, and activities observed in the synthetic dataset are shown in Figure 7(a-c). The learned actions (i.e.shown in Figure 7.a) correctly represent the four lanes formed on the side of the road. The learned motion patterns for synthetic dataset are shown in Figure 7(b). These motion patterns represent the four legal crossings and turnings. These motion patterns were learned from the action dependency matrix (as shown in Figure 7.b) returned by RTM. The $\{K_a \times K_a\}$ action dependency matrix should be interpreted as rows indicating action, and the brightness of each column represents dependency between actions. This dependency was used for motion pattern generation. The order in which motion patterns occur is discovered using the transition matrix. The activities discovered using the transition matrix is shown in Figure 7(c).Therefore, learned actions and their relationships correctly represent the motion patterns occurring at the intersection, i.e., horizontal and vertical crossing and turns (Figure 7.b), while the learned activity shown in Figure 6(c) correctly matches with the sequence shown in Figure 7(c).

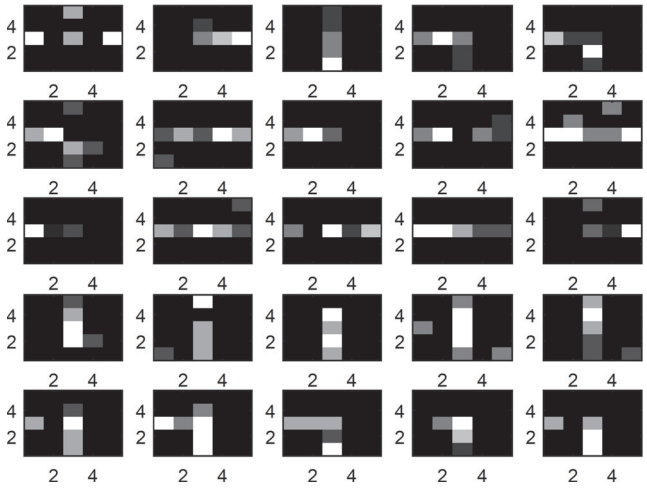


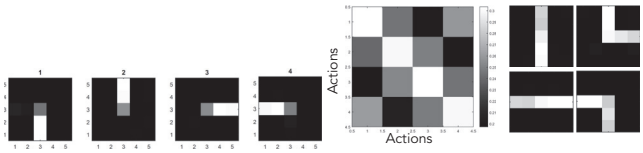
Figure 5: Synthetic dataset representing traffic observed at a 4-way traffic intersection.

(a) Lanes observed in the synthetic dataset.

(b) Traffic crossing present in the synthetic dataset

(c) Sequence of traffic crossing present in synthetic dataset.

Figure 6: Lanes, crossing and sequence of crossing observed in the synthetic dataset.



(a) Actions generated by RTM.

(b) Motion patterns generated using action-dependency matrix.

(c) Activities discovered by Activities Transition Matrix.

Figure 7: Actions, motion patterns and activities generated for Synthetic dataset.

3.2 Real dataset

We performed experiments on four video datasets. These datasets are captured by a single static camera. The videos mainly consists of the traffic flows that are dictated by traffic lights at either a four-way traffic intersection or traffic at roundabout. Following are the four datasets:

- QMUL Street Intersection Dataset: This video has traffic observed at a four-way traffic intersection. Traffic is flowing in vertical and horizontal directions and the sequence in which these flows is occurring is governed by traffic lights. This dataset consists of 45 minutes of video with the frame rate of 25 fps. The frame size is $\{360 \times 288\}$.
- Pedestrian Crossing Dataset: This video consists of traffic from vehicles flowing in a vertical direction with pedestrians walking in a horizontal direction. These flows are again controlled by traffic lights. This dataset contains 45 minutes of video of 25 fps and frame size of $\{360 \times 288\}$.
- Roundabout Dataset: This video contains traffic regulated by traffic lights observed at a roundabout. Traffic is flowing in a vertical and a horizontal direction. The vertical traffic is flowing from both directions but only one-way for horizontal traffic, i.e., from left-to-right. This dataset contains 60 minutes of video with the frame rate of 25 fps and frame size of $\{360 \times 288\}$.

(a) Actions learned using RTM for QMUL Street Intersection dataset.

(b) Actions learned using RTM for QMUL Pedestrian-crossing dataset.

(c) Actions learned using RTM for QMUL Roundabout dataset.

(d) Actions learned using RTM for MIT traffic dataset.

Figure 8: Actions learned using RTM.

- MIT Traffic Dataset: This video contains a street view of sparser traffic flow as compare to QMUL street intersection. This dataset contains 90 minutes of $\{360 \times 288\}$ pixel frame with frame rate of 30 fps.

Pre-processing and feature-extraction We used a 5-minute video from each dataset. Each video was divided into 10 second video clips (Hospedales, Gong, and Xiang 2009; Kuettel et al. 2010). Then optical flow was computed to generate visual-events for our model. To reduce the computation time, frames were resized from $\{360 \times 288\}$ to $\{72 \times 90\}$ using bicubic interpolation.Next, the bag of visual-events (thresholded optical flow) was input to RTM to generate actions and activities. We ran RTM for 500 iterations. For each dataset, the number of actions is set to 6 (i.e.$K = 6$). Setting the number of actions $K = 6$ for every dataset provided us with ease of demonstration. Additionally, if K is smaller, than scene actions might be a combination of multiple actions and if K is larger, than actions are too small; therefore, scene information will be compromised. Also, Dirichlet hyperparameters were set at $\alpha = 0.25, \beta = 0.05, \eta = 0.5$ for all experiments,but these hyperparameters can be estimated while inferring action assignments using Gibbs Sampling (Griffiths and Steyvers 2004).

Actions to motion patterns: In Figure 8(a) we show actions learned for the QMUL street-intersection datasets. Each action has semantic meaning associated with it. For example, in Figure 8(a) we show actions associated with traffic moving from north-south, south-north with interleaving turns, east-west and west-east. In Figure 8(b) we show actions detected on the pedestrian-crossing dataset. In Figure 8(c) we show actions learned for the roundabout dataset.

Motion patterns are a combination of actions that were co-occurring in time (i.e., observed together in the video).

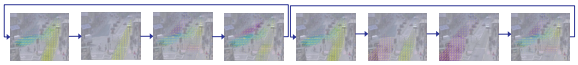
Our approach used dependencies between the actions to discover motion patterns instead of applying LDA again on the discovered actions by LDA. Additionally, LDA requires advance notification of how many motion patterns to be generated that restricts the generation of a complete set of motion patterns. In Figure 9 we have shown motion patterns discovered in the QMUL street-intersection dataset such as vertical traffic from opposite directions, vertical flow with left turn, and turns. For example, the vertical crossing motion pattern is a combination of vertical traffic from both directions and is shown in Figure 9(a). In the pedestrian crossing, Figure 9(b) shows the following motion patterns: vertical traffic flow in both directions, vertical turns, and pedestrians crossing from both horizontal directions. In the roundabout-intersection dataset, Figure 9(c) shows the following motion patterns: vertical bottom-up flow, horizontal flow, vertical right turn and flow, and vertical left flow. For the MIT traffic dataset the following motion patterns were discovered by the action dependency matrix: traffic from north-south and south-north, horizontal traffic from left taking left turn, and horizontal traffic from both sides with left turns. We show some of the MIT traffic motion patterns in Figure 9(d).

Identifying activities: Activities are sequences of motion patterns. The sequence identified by the Activity Transition Matrix for the street intersection as shown in Figure 9: vertical traffic from both direction is followed by vertical bottom traffic with turn, then vertical turns were followed by horizontal traffic from both sides one after the other, and back to the vertical flow. Activity for pedestrians crossing is shown in Figure 9(b), e.g., vertical traffic followed by pedestrians crossing alternatively. A sequence for the roundabout dataset is shown in Figure 9(c). The bottom traffic is followed by horizontal traffic. A sequence for the MIT traffic dataset is shown in Figure 9(d). Here, vertical traffic is followed by horizontal traffic.

4 Conclusion

We presented the Relational Topic Transition Model for discovering actions, motion patterns, and activities in a video. Our method is successful in discovering actions and activities in videos. We successfully applied our method on four real-video datasets and evaluated the method's validity on synthetic dataset as performed by (Hospedales, Gong, and Xiang 2012). To our knowledge, this is the first time RTM is being used for analyzing videos. In the future, we are planing to use tracklets as input.

(a) Activities for the QMUL street-intersection dataset. In this dataset usually vertical traffic is followed by turns and then horizontal flow.



(b) Activities for the pedestrian-crossing dataset. In this dataset traffic is flowing in vertical directions with interruption from pedestrians.

(c) Activities for the roundabout dataset. In this dataset traffic from bottom is followed from traffic from west direction.

(d) Activities for the MIT traffic dataset. In this dataset traffic is coming from north and south with interleaving right turn.

Figure 9: Activities for all the datasets.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Chang, J., and Blei, D. M. 2009. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, 81–88.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America* 101(Suppl 1):5228–5235.

Hospedales, T.; Gong, S.; and Xiang, T. 2009. A markov clustering topic model for mining behaviour in video. In *Computer Vision, 2009 IEEE 12th International Conference on*, 1165–1172. IEEE.

Hospedales, T.; Gong, S.; and Xiang, T. 2012. Video behaviour mining using a dynamic topic model. *International journal of computer vision* 98(3):303–323.

Kuettel, D.; Breitenstein, M. D.; Van Gool, L.; and Ferrari, V. 2010. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 1951–1958. IEEE.

Saleemi, I.; Shafique, K.; and Shah, M. 2009. Probabilistic modeling of scene dynamics for applications in visual surveillance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(8):1472–1485.

Stauffer, C. 2003. Estimating tracking sources and sinks. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 4, 35–35. IEEE.

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