Is that Scene Dangerous?: Transferring Knowledge Over a Video Stream

Omar U. Florez and Curtis Dyreson
Utah State University
{Omar.Florez, Curtis.Dyreson}@usu.edu

ABSTRACT

Activity mining in traffic scenes aims to automatically explain the complex interactions among moving objects recorded with a surveillance camera. Traditional machine learning algorithms generate a model and validate it with manually labeled data, which is a time-consuming and expensive task. The common issue is that these models often get outdated when external variables take place during posterior recording such as dynamic background, illumination, and different weather conditions. Those changes practically impose a new domain that often makes the original model inaccurate for clustering and classification tasks. If we directly apply a statistical model trained in one domain to other over the same stream, the performance of the algorithm will notably decrease due to distinct activity representations and different marginal and conditional distributions.

We approach this problem in two stages: 1) we present mature results on a hierarchical Bayesian model designed to represent every video scene as a multinomial distribution over topics. 2) we present early stage evidence of an algorithm to transfer knowledge across two instances of the hierarchical model described in the previous stage. A concrete example of this first stage consists of a simple (but efficient) algorithm to incrementally generate association rules to explain current traffic scenes as co-occurrence relationships between topics. This approach is especially useful when we do not have any labels in a target domain, but have some labeled information (*which frames contain dangerous scenes?*) in a source domain, by far the most frequent case in real surveillance systems. This algorithm clusters domain-dependent activities in the latent space and bridge them across domains via domain-independent activities. Our experiments show that our method is able to successfully compete with SVM to perform generalization when the temporal gap between source and target domain is large.

PIKM'12, November 2, 2012, Maui, Hawaii, USA.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information storage and retrieval; I.5 [**Pattern Recognition**]: Clustering

General Terms

Algorithms

Keywords

Scene understanding, Statistical learning, Video streams

1. INTRODUCTION

In many real-world applications, data takes the form of an ordered sequence of continuously arriving items called a *stream*. Over time, a huge amount of data can accumulate and the distribution of data within a stream can vary. Traditional modeling over a video stream includes the explicit representation of discrete elements and the frequency of their combinations in a single scan because of the continuous arrival of data.

Video can also be modeled as a data stream. Video is widely used in real-time monitoring applications, e.g., of an oil spill, a store entrance, or an airport. In traffic video streams, we are interested in discovering and monitoring the hidden rules that govern the behavior of multiple objects occurring in the same scene. Discovering these associations over different portions of the video streams raises three new issues, which go beyond traditional techniques.

1. Common behaviors describe *activities* - The similar patterns of continuous objects discovered in the stream (e.g., a car moving from right to left) need to be categorized under a discrete information unit called an *activity*.

2. No *a priori* knowledge of activities - The number of activities is not known in advance, rather they depend on the distribution of moving objects present in each video. Some method or model is needed to automatically infer activity information from a video.

3. Knowledge is *domain specific* - External parameters (dynamic background, illumination, different weather conditions, etc.) constantly impose new domains over the stream. Queries to an original model provide inconsistent results because of different activity and interaction parameters. We would like a method to reuse the knowledge acquired in other domains of the stream to response to novel queries successfully.

These challenging issues motivate our design of a framework for the analysis of traffic video streams. Our visual surveillance system is designed to automatically answer questions such as: "*Which is the most frequent scene seen so far?*", "*Is that scene dangerous?*", and "*Is this similar to what we learnt before?*". To do this every scene is modeled as a time window that contains a combination of zero or more activities made by individual moving objects. A time window circumscribes the interactions between activities found in a scene as documents containing a collection of words. We generalize this problem to study mechanisms to transfer knowledge acquired for a portion of the stream to other ones. Our goal is to reduce uncertainty for unlabeled data in a target domain by reusing the knowledge acquired in a prior machine learning model already trained. Rather than studying the concept drift for a particular model, we focus on bridging two unsupervised models through domain-independent activities to allow transfer of knowledge information (e.g., "dangerous scene") in the latent space. We submit this paper to the PIKM workshop in hope of a rich discussion on knowledge transfer and domain adaptation for automatic video surveillance.

1.1 Contributions

This paper makes the following contributions.

- We propose an unsupervised framework based on topic modeling that efficiently addresses the complete process of scene understanding over video data streams. Previous research (see Section 5) proposes algorithms that assumes a fixed number of activities in a video or same distribution of activities during learning and validation steps.
- We show practical evidence that this framework is able to explain complex activity relationships with simple co-occurrence rules.
- We propose an algorithm to perform knowledge transfer across different domains over a single video stream. To the best of our knowledge, ours is the first attempt to study this problem for activity mining in surveillance video streams.

1.2 Paper organization

The rest of this paper is organized as follows. Section 2 describes a method to discover activities from video data and explain the need of a hierarchy for this problem. Section 3 approaches the problem of transferring knowledge contained in labels across different sections of a continuous video stream. Section 4 shows ongoing experiments that demonstrate the usefulness of the proposed approach for real traffic video datasets. Section 5 discusses related work. Finally, Section 6 concludes the paper.

2. DISCOVERY OF ACTIVITIES

The problem of discovering activities in a video involves three kinds of information: *events*, *actions*, and *activities*. An *event* is an low-level interest point that represents a pixel with high variance in its spatio-temporal neighborhood. For a moving object, events occur in a bounding box forming a particular spatial arrangement of points that characterizes the *action* being performed. While a set of events characterizes an *action* (e.g., a car moving from right to left or a person walking in certain direction), *activities* are clusters of actions with similar event representation. This terminology and hierarchical relationship between *events*, *actions*, and *activities* have been adopted by the Computer Vision community, so we too use these common definitions. Given an input video, we take two consecutive frames and use a threshold to remove pixels with low intensity, as shown in Figure 1 (a). Then, we extract their events (gradient points) using a technique by Laptev et al. [5]. We evaluate connected components in Figure 1 (a) (represented as bounding boxes) to find moving objects in the scene as shown in Figure 1 (b). Finally, we place grids on those boxes to discretize the location of existing events into $n \times n$ small regions, as shown in Figure 1 (c). Note that we want that every connected component often corresponds to a single moving object, so we obtained better results by only considering rectangular bounding boxes, enclosing components, with width-to-length ratio in the interval of $(0.7, 1.3)$. When we divide the number of events found in every small region by the total number of events in a motion grid, we estimate the probability of finding an event in that region. For objects performing the same activity (G_1) in Figure 1 (c), we can see how the grids also show a similar spatial arrangement of events.

Our goal at this stage is to model how events are organized into activities. Thus, in this section we use a hierarchical model of two levels to generate activities in video data as multimodal probability distributions over events.The first level in the lower part of the hierarchy generates a mixture of events y_i that uniquely define an action with multimodal distribution θ_{ji}. The second level generates a list of activities G_0 distributed as the mixture model G_j over several multimodal distributions θ_{ji}. These two groups of information come from different, but related mixture models. The hierarchical way of forming activities seems to indicate that both groups share some mixture parameters. However, note that we do not know the number of mixture component in G_0 needed to represent the clustering process involved. In our case, it is difficult to specify *a priori* the number of event observations (regions in a grid) and activities needed to correctly interpret interactions in a traffic video. Our approach is to set the number of event observations as an external parameter dependent on the resolution of a particular video, but infer the number of activities by using a Dirichlet process in each group of actions. The use of a Dirichlet Process is justified by its property of providing a non-parametric estimation of the number of mixture components for groups of observations.

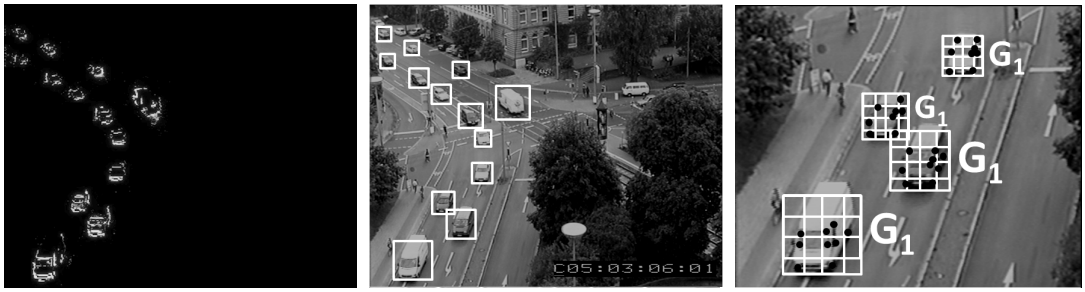
We first define the Dirichlet Process and then present a hierarchy of two Dirichlet Processes that can discover a number of activities in video data.

2.0.1 Dirichlet Process

Each event observation can be generated independently by a mixture component θ_{ji}. Let θ be a mixture component (cluster) associated to the event observation y_{ji}

DEFINITION 1 (DIRICHLET PROCESS). *A Dirichlet Process (DP) is a stochastic process that generates a distribution G in the form of an infinite mixture of components $\theta_i = \{\theta_1, \theta_2, ...\}$, a base distribution G_0, and a positive scaling parameter α.*

The construction of the Dirichlet Process can be formu-



(a) Segmentation (b) Bounding boxes (c) Motion grids and Events

Figure 1: **Describing activities of moving objects with events. (a) Motion information is segmented with pairs of consecutive frames (b) Bounding boxes enclose activity information (c) When zooming-in the frame, we can see how the spatial arrangement of events describes similar activities within 4×4 bounding boxes.**

lated with sequences of independent random variables $(\pi'_i)_{i=1}^{\infty}$ *and* $(\theta_i)_{i=1}^{\infty}$*, as originally stated in [8]:*

$$\pi'_i \mid \alpha, G_0 \sim Beta(1, \alpha)$$

$$\theta_i \mid \alpha, G_0 \sim G_0$$

such that the random distribution G *is then defined as:*

$$\pi_i = \pi'_i \prod_{l=1}^{i-1}(1 - \pi'_l)$$

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_k}$$

where δ_{θ_i} *is an atomic distribution centered on* θ_i*. For convenience, we shall abbreviate the construction of* π *as* $\pi \sim GEM(\alpha)$*. Note that* θ_i *is a multinomial probability distribution over event observations* y_i*. In other words, the random variable* θ_i *has a probability of being associated to the set of event* $y = \{y_1, y_2, ..., y_{n \times n}\}$*. Hence, the distribution base* G_0 *also needs to be distributed as a multinomial distribution. This property of having a family of multivariate probability distributions is especially found in the Dirichlet distribution*[1]*, so we model* G_0 *as being distributed as that distribution,* $G_0 \sim Dirichlet(D_0)$*.*

The Dirichlet Process generates a list of clusters of events $\theta = \{\theta_1, \theta_2, ...\}$ from the mixture model G that characterizes an activity based on the event observations y_i. Although this setting can represent appropriately one activity, it cannot represent several activities, which is needed for activity recognition in video data. The modeling of activities is defined as a hierarchy of two DPs that relates the generation and activities jointly.

2.1 The Hierarchical Model

We employ the Hierarchical Dirichlet Process (HDP) introduced by Teh et al. [10] to mutually learn both actions and activities by considering a second DP which models groups of actions θ_{ji} into activities G_j. The result is a hierarchical process which can be understood as the two level DP represented in Figure 2.

[1]This is the reason why a Dirichlet distribution is commonly denominated as a distribution of distributions.

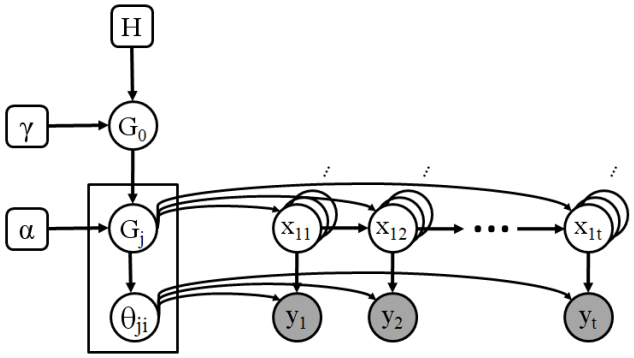


Figure 2: **A hierarchical process to find activities in video data. Each circle is a random variable and shading represents events observations from a grid.**

The lower level of the hierarchy generates an unbounded number of HMMs (Hidden Markov Models) that learn activities with an unknown number of states, considering event probabilities from a motion grid as observable variables. The upper level combines similar actions (learned in the HMM) into activities.

Lower Level.

The first level in the hierarchy constructs a variant of the Hidden Markov Model with state transitions distributed as G_j. The HMM is a doubly stochastic Markov chain in which a sequence of state variables $x = \{x_1, x_2, \ldots, x_T\}$ is hidden, but the sequence of observations $y = \{y_1, y_2, \ldots, y_T\}$ is observable. Changes between states are modeled with state transition probabilities and every state x_i is a multimodal variable that emits a discrete set of observations with some probability distribution. Traditionally, HMM assumes a Gaussian distribution for this property. However, it could represent even more complex observation behaviors when the output of the states is represented as the mixture of two or more Gaussians.

Every HMM is defined by the probability of each state to transit to other states and the probability of each state to emit an observation. In our model, both groups of information are assumed to be distributed as probability mixture models G_j for states and θ_{ji} for observations. A Dirichlet

Process is used to approximate each mixture model with an unknown number of mixture components. Since we do not assume an arbitrary number of states, the transition to an infinite number of states is modeled using a DP following the construction procedure presented in Definition 1.

$$G_0 \mid \gamma, H \sim GEM(\gamma)$$

$$G_j \mid \alpha, G_0 \sim DP(\alpha, G_0)$$

$$\theta_{ji} \mid \alpha, G \sim DP(\alpha_0, G_j)$$

for each $j = 1, 2, ...$, the probability θ_{ji} related to the activity j are learned with a HMM of states x and observations y, which have the following distributions.

$$\begin{array}{ll} x_t \mid x_{t-1}, (G_{ji})_{j=1}^{\infty} \sim G_{x_{t-1}} & \text{for states} \\ y_t \mid x_t, (\theta_{ji})_{j=1}^{\infty} \sim F(\theta_{x_t}) & \text{for observations} \end{array}$$

here, G_j is the distribution for the squared matrix that represents the transitions between states for the activity j. Different activities will be learned by HMMs with different distributions G_j.

Upper Level.

While the lower level generates a list of HMMs that recognizes individual activities, the upper level in the hierarchy selects the optimal HMMs associated to the activity j. The result is a list of activities $G = \{G_1, G_2, ...\}$ distributed as a mixture model G_0 with base distribution H, and a positive scaling parameter γ.

$$G_0 \mid \gamma, H \sim GEM(\gamma)$$

In other words, the base distribution G_0 generates the distributions G_j by grouping similar HMMs that learns similar event distributions θ_{ji}. Teh et al. [10] also use Gibbs sampling schemes to do inference under the HDP model. To detect the activity associated to a bounding box with a sequence of events observations $\{y_1, y_2, ..., y_{n\times n}\}$, first the trained HMM with highest log-likelihood score is selected. Second, the activity of the corresponding G_j associated to the item j is chosen.

3. TRANSFERING LABELS ACROSS DOMAINS

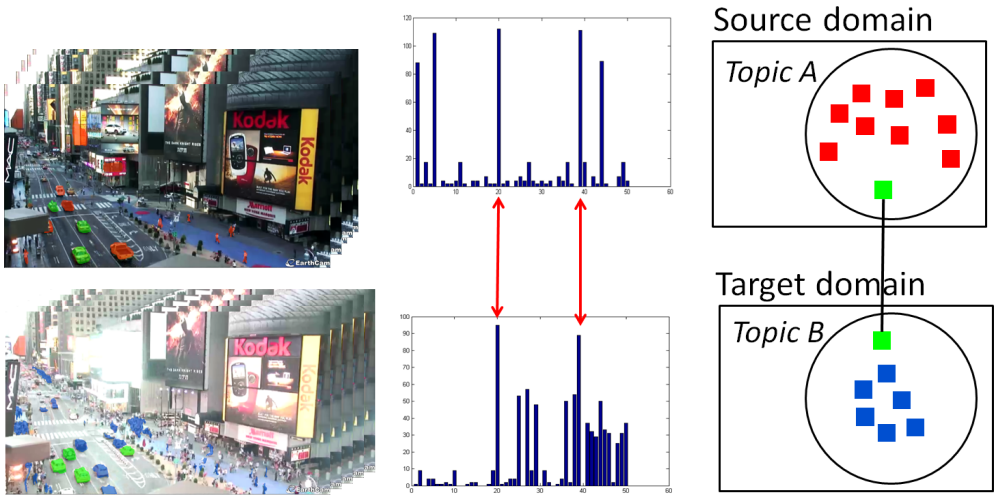
The mining of activities considering multiple domains exhibits different characteristics than traditional algorithms that also model activities as a combination of topics, but in one single domain. One big advantage of topic models is that they reduce the feature dimensionality (number of activities) into a discrete collection of topics. When considered together, topics provide a structure for a large collection of discrete data. Although topics are well-defined for a domain, they get partially preserved when we project them onto new domains. In our video problem, this happens because both source and target domains have different vocabularies (set of activities) and different marginal and conditional probabilities. These issues violate fundamental assumptions of topic modeling for mining crowd activities in video data. To further understand this point, Figure 3 shows two domains over the same traffic video stream. Figure 3 (a) corresponds to 2 hours of capturing scenes from a public web camera at different times (6AM vs. 6PM) on Lenon Street in New York[2]. Activities of moving objects are highlighted with colors. We choose morning and evening recordings in hope of finding different domains due to changes of illumination and volume of activities. Figure 3 (b) shows a histogram of 50 bins after projecting every event-based representation of moving objects (cf. Section 2) with the same hash function chosen at random with LSH in each domain. We do this to roughly count individual activities in the videos while providing comparable representations for this example. Note that some entries in the vocabulary are frequent in both domains. Their corresponding activities are highlighted in green in 3 (a) and corresponds to domain-independent activities. In other words, two different domains are transferable over the same stream if they share a set of activities that close the gap at the latent topic level. Figure 3 (c) shows how both the source and target domains are composed of domain-independent (in green) and domain-specific (in blue and red) activities.

What makes specially challenging the reuse of existing knowledge from a source domain is the lack of labeled training samples in the new domain. Topic labeling provides a conceptual summary in both domains as the activities being assigned to the same topic are positively correlated in individual time windows. Because of the subjectivity involved in detecting what is dangerous (not necessarily abnormal) in a collection of scenes, human labeling is used to improve precision and recall. The generation of a ground truth to detect dangerous situations is very expensive yet. Despite of crowd sourcing techniques to manually label tons of repetitive tasks such as Mechanical Turk [9]. The amount of time and money devoted to train a classifier in this way is prohibitively expensive and hinder its application in other domains. Based on these observations, we propose a method to transfer knowledge (i.e., scene labels assigned by human tagging) between two statistical topics models. Our algorithm is a generative topic model that differentiates relevant topics across different domains. However, not all topics and activities are relevant for cross domain representation. In fact, in our experiments less than 35% of all possible activities across domains is domain-independent. We present this algorithm in the next subsection.

3.1 Cross-Domain Knowledge Transfer

The algorithm to transfer labels containing knowledge across domains begins by finding topics in each domain with the the Hierarchical Dirichlet Process (HDP) explained in Section 2.1. Given N topics in the source domain θ^i and M topics in the target domain θ^j, we compute the *transference power* across domains by first counting the number of common activities between both domains ($\alpha \leftarrow \frac{\theta^i \cap \theta^j}{\theta^i \cup \theta^j}$) and then computing the Cross-Domain Transfer Coefficient (CDTC) by weighting with α the likelihood of every activity A of being generated by its corresponding topic, $P(A|\theta)$, regarding to domain-independent ($A \in \theta_{DI}$) or domain-dependent ($A \in \theta_{DD}$) groups of activities:

[2]Date: 6/15/2012, http://www.earthcam.com/usa/newyork/timessquare/?cam=lennon_hd



(a) Lenon Street (6AM and 6PM) (b) Vocabulary (c) Source and Target domains

Figure 3: **A motivating example for knowledge transfer in traffic video streams.**

$$CDTM \leftarrow \alpha \sum_{a=1}^{|\theta^i_{DI}|} p(A_a|\theta^i) + (1-\alpha) \sum_{k=1}^{2} \sum_{a=1}^{|\theta^i_{DP} \cup \theta^j_{DP}|} p(A_a|\theta^k)$$

A value of $\alpha > 0.5$ will indicate a more reliable knowledge transfer via domain-independent activities. Finally, for every topic in the target domain θ^j, we assign the label from the source domain θ^i with largest CDTM value, as described in Algorithm 1.

3.2 Why not use this for different streams?

The position of a traffic camera imposes a particular logic for the activity recognition. This is because different traffic lights and intersections will lead to obtain a different feature representation that is not comparable across different videos streams. That is why the current practice in crowd analysis is to use one camera to make inference and if we use multiple cameras, we should transfer information in a camera basis only. Otherwise, we could deal with the similar problem of compare topics extracted from different alphabets such as English and Chinese text documents.

4. EXPERIMENTS

In this section, we test the performance of our techniques with outdoor videos where moving objects describe traffic scenes governed by the state of multiple semaphores. The co-occurring interactions are modeled by frequent sets of activities with large confidence values over the video stream. We experiment on the following datasets: ***Street Intersection***[3] (*normal quality, 25fps, 90 minutes, 5 semaphores*), ***Karl-Wilhelm & Strabe Streets***[4] (*normal definition, 25fps, 2 hour, 3 semaphores*), and ***Roundabout Junction***[5] (*normal quality, 25fps, 2 hour, 3 semaphores*).

[3]http://www.eecs.qmul.ac.uk/∼jianli/Junction.html
[4]http://i21www.ira.uka.de/image_sequences/
[5]http://www.eecs.qmul.ac.uk/∼jianli/Roundabout.html

Algorithm 1 KnowledgeTransfer($Domain_1, Domain_2$)

1: //Output: $Topic_{source}$ more interesting for target domain
2: $Topic_{source} \leftarrow LDA(Domain_1)$;
3: $Topic_{target} \leftarrow LDA(Domain_2)$;
4: $N \leftarrow |Topic_{source}|$; //# Topics in source domain
5: $M \leftarrow |Topic_{target}|$; //# Topics in target domain
6: $TM[N][M] \leftarrow 0$; //Topic Matrix
7: **for** $i = 0$ to $N-1$ **do**
8: **for** $j = 0$ to $M-1$ **do**
9: $\theta^i \leftarrow Topic^i_{source}$;
10: $\theta^j \leftarrow Topic^j_{target}$;
11: //find number of common words
12: //find a specific weight for the transfer
13: $\alpha \leftarrow \frac{\theta^i \cap \theta^j}{\theta^i \cup \theta^j}$;
14: //CDTC: Cross-Domain Transfer Coefficient
15: $CDTM[i][j] \leftarrow 2\alpha \sum_{a=1}^{|\theta^i_{DI}|} p(A_a|\theta^i) +$
16: $(1-\alpha) \sum_{k=1}^{2} \sum_{a=1}^{|\theta^i_{DD} \cup \theta^j_{DD}|} p(A_a|\theta^k)$;
17: **end for**
18: **end for**
19: //find best transfer from the N source topics
20: **for** $j = 0$ to $M-1$ **do**
21: //find the highest CDTC value in the source domain for each target topic
22: $[log_likelihood, i] \leftarrow max(CDTC(:, j))$;
23: $Topic^j_{target}.label \leftarrow \theta^i_{source}$.label;
24: **end for**

Table 1: **Information on the training stage for each dataset.**

Dataset	Activities	Time to Compute
Street Intersection	37	4.38 hours
Karl-Wilhelm & Strabe	31	3.51 hours
Roundabout Junction	24	2.57 hours

The experiments are run on a 3.6 GHz Pentium 4 with 2 GB RAM and all the above datasets are publicly available to facilitate later experimental comparisons.

4.1 Experiment 1: Discovering Interactions

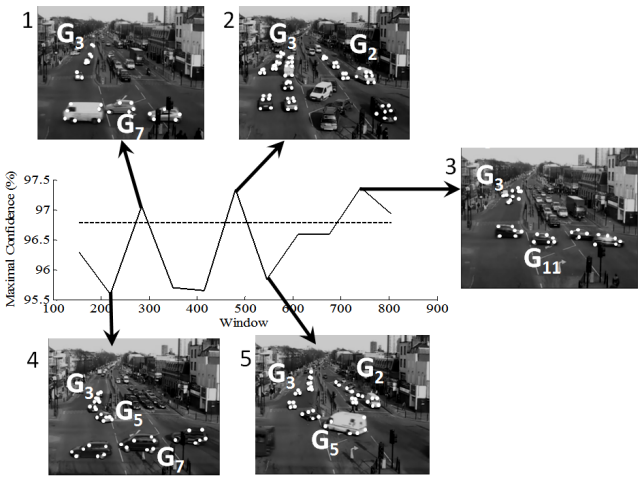
In this experiment, we study the significance of the generated rules to understand the co-occurrence dependencies between activities. Time windows of size $/w/ = 25$ frames is a value that works in all the datasets in order to find activities temporally correlated in the same window. We consider the well-known Apriori algorithm to efficiently extract rules based on the set of topics discovered in each video.

The number of transactions, topics, and the processing time to discover association rules for every dataset are summarized in Table 3. The **Street intersection** dataset exhibits more topics than the **Karl-Wilhelm & Strabe** dataset since five traffic lights decomposes complex activities into a large number of well-defined scenes. On the other hand, the **Roundabout Junction** dataset contains a few number of topics due to the limited types of activities performed and considerable amount of frames with no activities. The processing time to generate rules seems to be proportional to the number of topics discovered in each dataset.

Table 2: **Information on the datasets preprocessed to discover association rules.**

Dataset	Time windows	Topics	Rule generation time
Street Intersection	$\sim$ 15000	37	11.52 min.
Karl-Wilhelm & Strabe	$\sim$ 11000	31	8.12 min.
Roundabout Junction	$\sim$ 7000	24	6.33 min.

We consider a minimum support value of 4%, a minimum confidence value of 90%, and activity clusters with more than 10 elements in order to generate representative rules. We thus extract rules from the **Street Intersection** (37 topics and 16 rules), **Karl-Wilhelm** & **Strabe** (31 topics and 13 rules), and **Roundabout Junction** (24 topics and 10 rules) datasets. In these datasets as more constraints govern the activities (e.g., traffic lights, one-way roads, intersections, etc.), more topics are generated and more frequent rules are discovered. This evidence seems to indicate that every constraint imposes an underlying logic that fragments complex activities into a large number of small scenes, which are easy to represent with events and form well-defined activities, and therefore are likely to be frequent during the video. For the **Street Intersection** dataset, some of the rules uncovered with the algorithm proposed in this paper



(a) Scenes

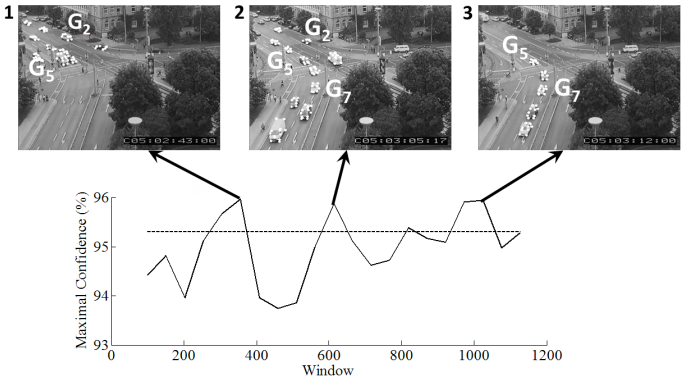
Scene	Rule	Confidence	Support
1	$\{G_7\} \rightarrow \{G_3\}$	97.2%	7.6%
2	$\{G_2\} \rightarrow \{G_3\}$	97.3%	11.2%
3	$\{G_{11}\} \rightarrow \{G_3\}$	97.4%	8.8%
4	$\{G_5\} \rightarrow \{G_3, G_7\}$	95.5%	6.5%
5	$\{G_3, G_5\} \rightarrow \{G_2\}$	95.6%	8.2%

(b) Rules

Figure 4: **Experiment on the Street Intersection dataset. (a) A selection of high confidence association rules. (b) Scenes of the Street Intersection dataset with high confidence values over time. Changes between scenes represent transitions between significant scenes.**

are depicted in Figure 4 (a) and detailed in Figure 4 (b). The first three rules are high-confidence associations that suggest a strong correlation between vehicles moving in parallel lanes (G_2 and G_3) or those moving from side to side (G_7 or G_{11}) while other vehicles move away from the center to the top left of the scene (G_3). Those activities are mutually exclusive since there are five traffic lights that prevent vehicles moving from side to side from colliding with those moving across the parallel lanes. Consider the first rule $\{G_7\} \rightarrow \{G_3\}$ as an example. The last two rules indicate the co-occurring dependency between cars turning right (G_5) while others that are moving from right to left (G_7). This behavior, exemplified by rule $\{G_5\} \rightarrow \{G_3, G_7\}$, is justified since those vehicles use the same traffic light to move from the bottom right part of the scene to either the bottom left or the top left edge, as seen in scene 4 and 5 of Figure 4.

For the **Karl-Wilhelm** & **Strabe** dataset, three confident interactions are shown in Figure 5 (a) and expressed with rules in Figure 5 (b). We notice the regular presence of the activity G_5 in those scenes. This behavior is reasonable since the activity G_5 corresponds to vehicles going along Strabe avenue, a very busy road in the dataset. The first rule $\{G_2\} \rightarrow \{G_5\}$ exemplifies the interaction of vehicles going in parallel lanes without restrictions. The second rule is similar, but additionally contains the activity of cars going from the center to the bottom left of the screen (G_7). Furthermore, the usual interaction of cars going straight in

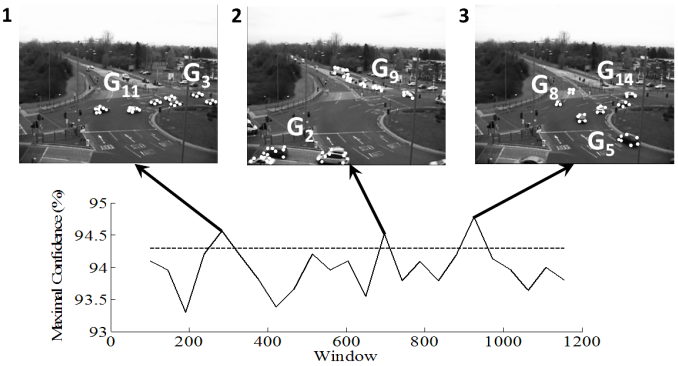


(a) Scenes

Scene	Rule	Confidence	Support
1	$\{G_2\} \rightarrow \{G_5\}$	95.9%	21.3%
2	$\{G_2, G_5\} \rightarrow \{G_7\}$	95.7%	17.8%
3	$\{G_5\} \rightarrow \{G_7\}$	95.6%	24.6%

(b) Rules

Figure 5: **Experiment on the Karl-Wilhelm & Strabe dataset. (a) A selection of high confidence association rules. (b) Scenes of the Street Intersection dataset with high confidence values over time.**



(a) Scenes

Scene	Rule	Confidence	Support
1	$\{G_{11}\} \rightarrow \{G_3\}$	94.5%	27.3%
2	$\{G_9\} \rightarrow \{G_2\}$	94.4%	13.5%
3	$\{G_5, G_{14}\} \rightarrow \{G_8\}$	94.8%	19.8%

(b) Rules

Figure 6: **Experiment on the Roundabout Junction dataset. (a) A selection of high confidence association rules. (b) Scenes of the Street Intersection dataset with high confidence values over time.**

the avenue (G_5) and then turning right after that (G_7) is explained by rule $\{G_5\} \rightarrow \{G_7\}$.

For the **Roundabout Junction** dataset, we show common interactions in Figure 6 (a) and detailed in Figure 6 (b). The roundabout in the video segments motion of vehicles into multiple activities. The first rules $\{G_{11}\} \rightarrow \{G_3\}$ represents the sequence of car activities going straight (G_3) and then joining the roundabout (G_3). By contrast, the second rule $\{G_9\} \rightarrow \{G_2\}$ explains the co-occurrent relationship of vehicles taking lanes separated by the roundabout. Finally, while some vehicles circulate alongside the roundabout emerging as two activities (G_5 and G_{14}), another set of cars take a different way by turning left from the center to the upper part of the scene (G_8). These sequences of co-occurring associations reflect the transitions between significant scenes in datasets governed by multiple lanes, traffic lights, and a roundabout.

4.2 Experiment 2: Knowledge Transfer

In this experiment we show our preliminary results on the transfer algorithm across different domains for each dataset. We generate a ground truth in every dataset with crowdsourcing by asking users to manually tag each each consecutive non-overlapping time windows of 10 seconds with either "'dangerous"' or "'non-dangerous"'. We used the first 20 minutes of each dataset for unsupervised learning in order to take last 20 minutes for testing. Our goal is to test the transferring between the 2 farthest domains of every video. Pairs of consecutive frames are processed to identify moving pixels, events, and connected components. The observations consist of bounding boxes around moving objects with resolutions of 8×8 for the **Karl-Wilhelm & Strabe** dataset and 4×4 for the **Street Intersection** and **Roundabout Junction** datasets. This is because the camera in the first dataset is placed on a far building, so we need grids with higher resolutions to describe small objects. This process provides a collection of unlabeled motion grids to the hierarchical model. We do not assume any prior knowledge in the number activities to be discovered. DP parameters were fixed at $\{\alpha = 11, \gamma = 0.9\}$.

Dataset	SVM	Transfer Knowledge
Street Intersection	83.41%	73.41%
Karl-Wilhelm & Strabe	68.81%	71.15%
Roundabout Junction	62.17%	75.82%

Table 3: Average precision for scene recognition.

Table 3 shows the average precision of the transfer algorithm to correctly detect the label of the frames in the target domain. To make a competitive comparison, we train a supervised algorithm SVM [3] with Gaussian kernel and give it the time windows coming from the first 20 minutes to then perform inference in the target domain. The feature vector for SVM is formed by the number of activities found for each topic in each domain. Note that in this case, we use for the target domain, the same topics discovered with HDP in its corresponding source domain to enable the same vocabulary of activities for both in training and testing. The results are encouraging as TransferKnowledge outperforms SVM when the gap between domains is large (e.g., 80+ mins. for the Karl-Wilhelm & Strabe and Roundabout Junction datasets). This is because SVM constructs hyperplanes based on features that are less consistent as larger is the gap between domains. This problem is even exacerbated when new interactions occur and HDP do not correctly group new activities into existing topics. However, when the gap is shorter (e.g., 50+ min for the Street Intersection dataset) SVM outperforms our model as topics are still stable enough to perform good generalization.

5. RELATED WORK

In this section we compare our approach with related efforts. For clarity, we keep our comparison focused in each stage of the video process (i.e., discovery of activities and knowledge transfer for surveillance videos).

The recognition of activities in video data is an open problem that has received much attention lately. Commonly, low-level visual features and actions have been modeled and classified to provide interpretation of activities. While the traditional way to categorize existing research is by motion representation such as local features (e.g., changes in velocity, changes in curvature of motion trajectories, and gradients) or global features (e.g., key frames), recent research has employed hierarchical Bayesian models such as LDA [2] and HDP [10] to cluster local motions into activities successfully, c.f., [4, 11, 12]. The above research has led to techniques that can discover atomic activities, but such techniques omit the complex interactions between activities commonly present in video data. Wang et al. [11] approach this problem by adding one more level to the hierarchy of the LDA and HDP Bayesian models and providing extended versions of integral probabilistic hierarchical Bayesian models (LDA, HDP, and Dual-HDP mixture models) to cluster moving pixels into atomic activities and interactions. Similarly, Li et at. [6] infer global behavior patterns through modeling behavior correlations through a hierarchical probabilistic Latent Semantic Analysis (pLSA). Both techniques, however, learn global interactions disregarding temporal information. By contrast, our on-line technique relates frequent activities in a transaction and removes those that become infrequent over time. In other words, by decoupling both the discovery of activities and interactions, we can incrementally learn interactions without assuming the same probability of co-occurring relationships over time, a reasonable scenario imposed by the processing of continuous video streams.

Despite of all the existing research in activity mining in surveillance videos, little has been done on analyzing collaborations across topics generated in two domains. Consider the case of sudden changes on background, different weather conditions, and variation on illuminations. In most of these cases, the distribution of activities in future frames is different. It is often hard for researchers to establish such cross videos interactions to recognize dangerous activities from normal ones. Cross-domain learning often exhibit very different challenges compared to traditional activity models in the same domain. For instance, Jain et al. [1] consider a variation of LDA, People LDA, to connect words to face images. Here the words act like labels which are easy to count and evaluate co-occurrences values. In our case we focus on transferring learning from domains that are both coming from videos. More recently, Li at al. [7]. proposed a similar system than PeopleLDA, but considering the transfer of knowledge between textual features and large video data.

6. CONCLUSION

In this paper, we propose a framework to find approximate co-occurring associations from video stream data considering unsupervised clustering of events (low-level visual features) into activities. We define activities as actions described by similar event distributions. A hierarchy of two stochastic processes is used to avoid considering an arbitrary number of activities in the video. The most visible aspects of this effort is the incremental generation of rules that discover the interaction of frequent activities for current scenes and the ability to make generalization over different domains of the same video stream. Our experimental results show that our approach efficiently and automatically discovers and transfer sets of activities in a video stream while evaluating their frequent occurrence and co-occurring relationships.

ACKNOWLEDGEMENTS

Partial funding for this research was provided by NSF EAGER (Award Number 1144404) and by IBM (Scalable Data Analysis for a Smarter Planet Innovations Award).

7. REFERENCES

[1] A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

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

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[4] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *International Conference on Computer Vision*, pages 1165–1172, 2009.

[5] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, June 2008.

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

[7] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Video summarization via transferrable structured learning. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 287–296, New York, NY, USA, 2011. ACM.

[8] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 2(476):639 – 650, 1994.

[9] E. Simperl. Crowdsourcing semantic data management: challenges and opportunities. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, pages 1:1–1:3, New York, NY, USA, 2012. ACM.

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

[11] X. Wang, X. Ma, and W. 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.

[12] J. Yin and Y. Meng. Human activity recognition in video using a hierarchical probabilistic latent model. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–20, 2010.