Unusual Event Detection via Multi-camera Video Mining

ICPR 2006 submission

Abstract

This paper describes a framework for detecting unusual events in surveillance videos. Most surveillance systems consist of multiple video streams, but traditional event detection systems treat individual video streams independently or combine them in the feature extraction level through geometric reconstruction. Our framework combines multiple video streams in the inference level, with a coupled hidden Markov Model (CHMM). We use two-stage training to bootstrap a set of usual events, and train a CHMM over the set. By thresholding the likelihood of a test segment being generated by the model, we build a unusual event detector.

We evaluate the performance of our detector through qualitative and quantitative experiments on two sets of real world videos.

1. Introduction

In response to the heightened demand for security, more and more video surveillance systems are being deployed. However, the number of trained personnel to watch these systems is too limited to analyze the vast amounts of captured video. From the London subway bombing attack [11] to the recent Bangladesh bombing attack [10], there were suspicious activities before the attack. Individually considered, each event does not exhibit sufficient significance to trigger an alarm, but if we combine the evidence from all relevant sites, the unusual events become more obvious. To detect events of such distributed nature, we propose a framework for unusual event detection using multiple surveillance video streams.

We first define the term "unusual". Unusual events that are of interest for security should meet one of the following two criteria: 1) It seldom occurs. 2) It is previously unseen. The rarity of unusual events limits available training data for such events. Furthermore, the distinctions between two unusual events can be as large as those between unusual events and usual events. Therefore it is not feasible to train a general model for the unusual events. On the other hand, we have abundant training data for usual events, and usual events tend to cluster well. Thus we propose training a model for usual events, and detecting unusual events that deviate from the usual event model.

Our work is mostly related to the semi-supervised unusual event detection framework proposed by Zhang *et al.* in [21]. They used MAP adaptation to detect unusual events in a semi-supervised fashion, as we described in detail in Section 2. They showed that the performance of the semi-supervised approach depends on the number of iterations.

Our approach alleviates this difficulty with a two-stage training process. In the first stage, we generate a training set of usual events by combining unsupervised clustering with user feedback. In the second stage, we train a more precise model for the usual events using the training set. An overview of the training process is shown in Figure 1.

2. Related Work

The problem of event detection and recognition is generally formulated as classification of time series data. Supervised approached learn a model of the reference event sequence from training samples and devise a matching criterion to both accommodate variations within one event class and discriminate between different event classes. Dynamic Time Warping (DTW), a matching technique widely used for speech recognition, has been used in recognizing human motion patterns [2]. Finite-state machines (FSM) are used for modeling vehicle motion from aerial images [7]. HMMs generally out-perform DTW on undivided time series. Brand *et al.* [4] applied entropy minimization to regulate the complexity of HMMs, and used the trained HMMs to detect unusual behavior in office activities as well as to monitor traffic. Oliver *et al.* [17] compared HMMs with CHMMs for modeling interactions like following and meeting, and showed that CHMMs are more efficient and accurate. Our CHMM differs from theirs in that each chain of our CHMM corresponds to one video stream, while on their CHMM, each chain corresponds to a 2D blob of one person from a single static camera.

One common assumption of the above approaches is that the events of interest are known in advance and can be modeled accordingly [8]. To detect unusual events which are not seen before, researchers have proposed unsupervised and semi-supervised approaches. Hongeng [14] used a Markov network to cluster and infer events in a dinner table setting scene. Zhong and Shi [22] proposed a SVD based feature

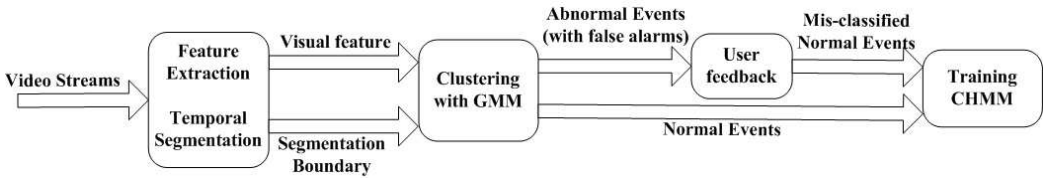

Figure 1. Overview of the training process

selection method for unsupervised unusual event detection. Zhang *et al.* [21] used a semi-supervised approach to train models for both usual and unusual events. They start from an ergodic hidden Markov model (HMM) for usual events. If a test event does not fit the model, they classify it as unusual and branch the usual event model to refit the usual event.

More recently, Boiman and Irani [3] proposed to model the set of usual events as an ensemble of spatial-temporal image patches, and detect irregularity in a test video by evaluating the similarity between the test ensemble with the training ensemble.

The novelty of our approach is that we combine multiple video streams with CHMM and applied it to semi-supervised unusual event detection.

3. Preprocessing: Visual Feature Extraction and Event Segmentation

Given an input video stream, we use adaptive background subtraction [20] to find the foreground region, and group the regions into connected components, and fit a bounding box around each connected component. The i^{th} bounding box is described with a 4D vector B_i = [x, y, w, h], where x and y are the image coordinates of the center of the bounding box; w and h are the width and height of the bounding box. The subscript i are sorted in descending order of the size of the bounding boxes.

Assuming there are C video streams captured at the scene, we select one stream as the reference stream for segmentation, and define an event as a sequence of frames where at least one blob is detected in the reference stream. If there is no motion blob for several consecutive frames in the reference stream, we determine it is a segmentation boundary between two events.

The flowchart of visual feature extraction and event segmentation is shown in figure 2.

4. Stage 1: Mining Events by Unsupervised Clustering

Since it is tedious to pick out unusual events from a large number of unlabeled events, we first use an unsupervised method to find the clusters in the unlabeled set, and label the events in the largest clusters as usual events.

4.1. Hierarchical Clustering with Approximate KL-divergence

To cluster the video segments, we first extract a signature for each event segment by fitting a Gaussian Mixture Model (GMM) distribution over all the bounding boxes detected in the frames belonging to that event segment. The signature consists of the model parameters of the GMM (the mean, covariance and weight of each Gaussian kernel).

Given the signatures, we use agglomerative hierarchical clustering to group similar events. The similarity between two groups of signatures are defined as the approximate KL-divergence between the two GMMs [12]. The GMM for each group consists of all the Gaussian kernels in the signatures in that group, with their weights normalized according to the total number of events in each group.

4.2. User Feedback to Fine-tune the Set of Usual Events

Based on the clustering result, all the events in the large clusters are labeled as usual. Since some of the usual events may be scattered in the small clusters, we examine the events in the small clusters[1] to pick out the usual events and put them into the set of usual events. This will help fine-tune the boundary between usual and unusual events and provide a way to impose semantic meaning upon the usual events.

5. Stage 2: CHMM-based Unusual Event Detection

We train a coupled hidden Markov model (CHMM) [5][6] over the set of usual events. Given an observed test video segment, we evaluate the likelihood that the segment is being generated from the CHMM. If the likelihood is below a threshold, the segment is classified as unusual.

5.1. Combine Multiple Video Streams with Coupled Hidden Markov Model

CHMM's are widely used to analyze multiple streams of intrinsically related data flows. We choose the coupling structure shown in Figure 3 in our CHMM.

In Figure 3, each node represents a discrete hidden variable $X_i[k]$, where $i = 1 \ldots C$ is the chain index, C is the

[1] Such events consist only a small fraction of the training set.

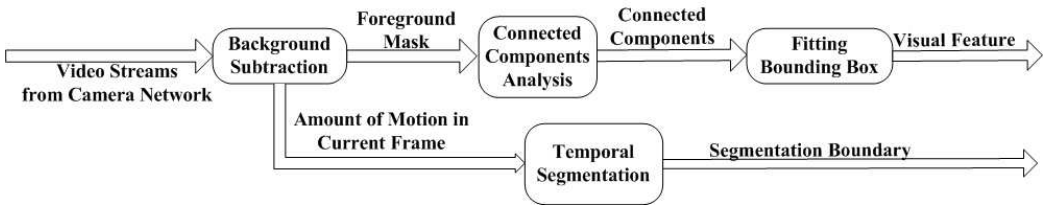

Figure 2. Flowchart of visual feature extraction and event segmentation

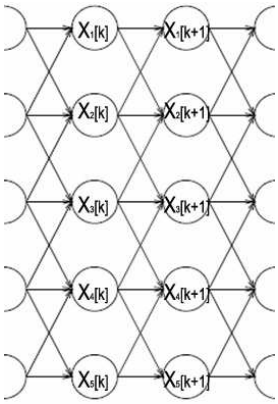

Figure 3. The structure of our coupled hidden Markov

number of chains and k is the time index. Each chain corresponds to one camera. Chain X_i is coupled with the neighboring Chain X_{i-1} and Chain X_{i+1}.

Meanwhile, each hidden variable $X_i[k]$ is associated with the observable variable $Y_i[k]$, which is defined as a 12 D vector $Y = [B_1, B_2, B_3]$ consisting of the location and size of the three largest bounding boxes detected in Frame k from Stream i. When the number of motion blobs is smaller than three, we pad zeros to the corresponding entries[2].

To parameterize the probability density functions (PDF) of the observation vectors, we again use GMM distribution. We assume the 12 components in the vector are independent and thus the covariance matrix for each Gaussian kernel in the GMM distribution is diagonal.

For exact inference in the CHMM, we use the frontier algorithm [23], which is an extension of forwards-backwards algorithm, to recursively compute the posterior probability $P(X_i[k]|Y_i[1]\dots Y_i[k])$, for each chain $i = 1\dots C$. The basic idea of the forwards-backwards algorithm is to exploit the d-separation [15] between the past and the future by the hidden variable $X[k]$. In CHMM, the set of all hidden nodes at one time slice, i.e. $X_i[k]$, $(i = 1\dots C)$, d-separate the past from the future. The frontier algorithm "sweeps" a Markov blanket across the Dynamic Bayesian Network (DBN), first forwards and then backwards. The nodes in the Markov blanket are called the "frontier set". In the case of the CHMM shown in Figure 3, the complexity of the frontier algorithm is $O(TN^{D+1})$, where T is the total number of time slices, and N is the number of states for each hidden node. In particular, we use the junction tree implementation for 2 time-slice temporal Bayesian Networks [16].

For learning the model parameters, we use the Expectation Maximization [1] algorithm. The E-step is done in the same way as exact inference. The M-step is done with stochastic sampling.

6. Experimental Results

To evaluate the performance of our framework, we collected 7 days (56 hours) of synchronized video streams from 5 cameras mounted in a mail-room. A schematic view of the mail-room with the cameras appears in Figure 4.

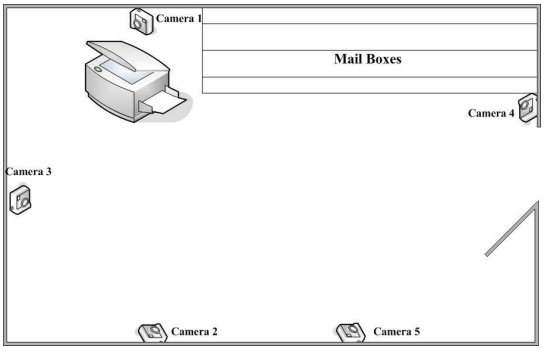

Figure 4. A schematic view of the mail-room with the cameras

We choose the video stream from Camera 3 as the reference stream for event segmentation, because it covers the entrance to the mailroom. From 56 hours of video, we extracted 408 segments ranging from 2 to 60 seconds. At Stage 1 of the training process, 397 segments were clustered in the largest cluster and automatically labeled as usual. We use them to train a CHMM. There are two major parameters of the CHMM: one is the number of states that the hidden variable X can take, the other is the number of Gaussian kernels in the PDF of the observable variable Y. We tried CHMM's of 3 to 8 states, with observation PDF of 2 to 9 kernels. Using the Minimum Description Length Principle (MDL) [13], we select the best model, which is the 4-

[2]One drawback of padding zeros is that it introduces many observations on the axes, or at the origin of the feature space, with very little variances. To deal with this problem, we are still investigating computationally feasible alternatives to model sequences of variable length feature vectors.

state CHMM with 6-kernel GMMs as the observation PDF. It reaches a balance between model complexity and overall likelihood of generating the training set.

By thresholding the likelihood of the 408 events given the CHMM, we detected 21 unusual events. Some snapshots of the unusual events in our mail-room videos are shown in Figure 5. They correspond to such events as someone fumbling with more than one person's mailboxes, someone opening the copy machine, or someone trying to tamper with the surveillance cameras.

After examining those classified as usual events, we found that they correspond to instances of people walking by the entrance of the mail-room, picking up printouts and picking up mail from one mailbox. Some snapshots of the usual events are shown in Figure 6.

Because of space constraints, we only show one camera view for each set. Figure 6 shows images captured from Camera 3, while Figure 5 shows images captured from Camera 4.

To evaluate the performance of our algorithm quantitatively, we use the labeled data from the Terrascope Dataset [9]. The dataset consists of synchronized videos captured by nine different cameras, deployed over several different rooms and a hallway in a office setting. The videos were recorded under the following four scenarios:

1. Group meeting,
2. Group exit and intruder,
3. Suspicious behanvior/theft, and
4. Natural video sequences.

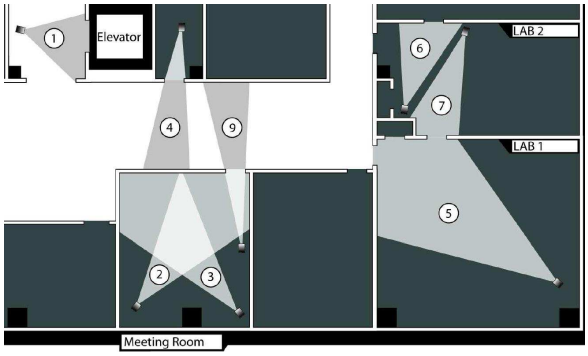

Figure 7. A schematic view of the collection space of the Terrascope Dataset, courtesy of Christopher Jaynes, University of Kentucky

Figure 7 shows a schematic view of the collection space of the Terrascope Dataset. We used five video streams from Camera 2, 3, 4, 5 and 9 in Figure 7. The stream captured by Camera 2 is chosen as the reference stream for event segmentation. We extracted 43 segments of video events from all the four scenarios. The ground truth labels are assigned according to the scenario. Event segments extracted from Scenario 1) and 4) are labeled as usual events, while those from Scenario 2) and 3) are labeled as unusual.

At Stage 1 of the training process, 38 segments were clustered in the largest cluster and automatically labeled as usual. To maximize the objectivity of the result, we bypassed the user feedback step described in Section 4.2, and directly use the 38 segments to train a 4-state CHMM with 6-kernel GMMs as the observation PDF.

We then rank the events according to their likelihood of being generated by the trained CHMM. By cutting of the ranked list from different points, we classified and events into two sets: unusual and usual. We generate a Receiver Operating Characteristic (ROC) curve for the unusual event detector. We compare our CHMM-based detection detector with a HMM-based detector, where we assume the video streams are independent and train a separate HMM [18] for each video. Figure 8 shows the ROC curves of the two detectors.

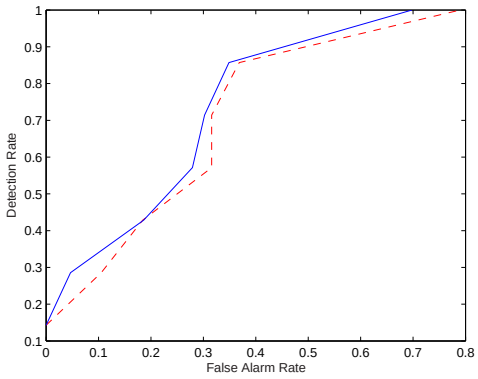

Figure 8. Comparison between ROC Curves. The blue solid line is the ROC curve of the CHMM-based detector and the red dashed line is that of the HMM-based detector.

We found that CHMM-based detector out-performs the HMM-based detector, especially at the lower false alarm region. We also notice that the ROC curves are not convex, and Scott [19] showed that a non-convex classifier can be improved because it is always possible to operate the classifier on the convex hull of its ROC curve, with a randomized rule that combines two operating points. Therefore, both detectors can be improved by introducing a randomized rule.

7. Conclusions

In this paper, we proposed a multi-camera mining framework for unusual event detection in surveillance video. We use two-stage training to bootstrap a probabilistic model for the usual events, and detect unusual event by thresholding

Figure 5. Snapshots of the unusual events, from the view of Camera 3

Figure 6. Snapshots of the usual events, from the view of Camera 4

the likelihood of a test event being generated by the usual event model.

The main contributions of our video mining method are the followings.

1. We combine evidence from multiple video streams with a coupled hidden Markov model (CHMM). Our experimental results show that it outperforms the HMM-based detector.

2. We proposed a two-stage training framework to bootstrap the usual event model. It eliminates the need for manually labeling the set of usual events.

Our framework has been evaluated qualitatively on 56 hours of video collected by the camera network installed in the mail-room of our lab, and quantitatively on the publicly available Terrascope Dataset [9].

In the future we are going to concentrate on the following two directions.

1. Identify the motion blobs by their appearance and location throughout the video. By integrating the identity into the visual feature, we can improve the spatial alignment in observable nodes of the CHMM. Another related issue has to do with the observation vectors that describe the observable nodes. Current solution of padding zeros is straightforward but can be improved with a more systematic model for sequences of variable length observations.

2. Select the model complexity adaptively. Based on the processing capacity of the end user, we can regulate the number of unusual events to be reported by adapting the model complexity.

References

[1] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models, 1997. 3

[2] A. F. Bobick and A. D.Wilson. A state-based technique to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1325–1337, Dec. 1997. 1

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *Proc. IEEE International Conference on Computer Vision*, pages 1985–1988, Beijing, China, Oct. 15-21 2005. 2

[4] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):844–851, 2000. 1

[5] M. Brand and N. Oliver. Boltzmann chainhidden Markov models for complex action recognition. In *Advanced in Neural Information Processing Systems*, 1995. 2

[6] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 994–999, 1997. 2

[7] F. Bremond and G. Medioni. Scenario recognition in airborne video imagery. In *Proc. Int. Workshop Interpretation of Visual Motion*, pages 57–64, March 1998. 1

[8] H. Buxton. Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, 21(1):125–136, 2003. 1

[9] N. S. C. Jaynes, A. Kale and E. Grossmann. The terrascope dataset: A scripted multi-camera indoor video surveillance dataset with ground-truth. In *Proceedings of the IEEE Workshop on VS PETS*, October 2005. 4, 5

[10] CNN. Bangladesh bombs: 2 dead, 50 held, 2005. 1

[11] CNN. More than 50 dead in London attacks, 2005. 1

[12] J. Goldberger and H. Aronowitz. A distance measure between gmms based on the unscented transform and its application to speaker recognition. In *Interspeech*, pages 1985–1988, Lisbon, Portugal, September 4-8 2005. 2

[13] P. Grunwald. A tutorial introduction to the minimum description length principle. In P. Grunwald, I. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Application*, chapter 1-2. MIT Press, Cambridge, 2005. 3

[14] S. Hongeng. Unsupervised learning of multi-object event classes. In *Proc. British Machine Vision Conference*, 2004. 1

[15] F. V. Jensen, F. V. Jensen, and F. V. Jensen. *Introduction to Bayesian Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. 3

[16] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002. 3

[17] N. M. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000. 1

[18] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989. 4

[19] M. J. J. Scott, M. Niranjan, and R. W. Prager. Realisable classifiers: improving operating performance on variable cost problems. In *Proc. British Machine Vision Conference*, September 1998. 4

[20] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):747–757, 2000. 2

[21] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted hmms for unusual event detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, July 2005. 1, 2

[22] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 819–826, Washington DC, July 2004. 1

[23] G. Zweig. *Speech Recognition with Dynamic Bayesian Networks*. PhD thesis, University of California, Berkeley, 1998. 3