Semi-supervised Adapted HMMs for Unusual Event Detection

Dong Zhang[1,2], Daniel Gatica-Perez[1], Samy Bengio[1,2] and Iain McCowan[1] *
[1]IDIAP Research Institute, Martigny, Switzerland
[2]Swiss Federal Institute of Technology, Lausanne, Switzerland
{zhang, gatica, bengio, mccowan}@idiap.ch

Abstract

We address the problem of temporal unusual event detection. Unusual events are characterized by a number of features (rarity, unexpectedness, and relevance) that limit the application of traditional supervised model-based approaches. We propose a semi-supervised adapted Hidden Markov Model (HMM) framework, in which usual event models are first learned from a large amount of (commonly available) training data, while unusual event models are learned by Bayesian adaptation in an unsupervised manner. The proposed framework has an iterative structure, which adapts a new unusual event model at each iteration. We show that such a framework can address problems due to the scarcity of training data and the difficulty in pre-defining unusual events. Experiments on audio, visual, and audio-visual data streams illustrate its effectiveness, compared with both supervised and unsupervised baseline methods.

1 Introduction

In some event detection applications, events of interest occur over a relatively small proportion of the total time: e.g. alarm generation in surveillance systems, and extractive summarization of raw video events. The automatic detection of temporal events that are relevant, but whose occurrence rate is either expected to be very low or cannot be anticipated at all, constitutes a problem which has recently attracted attention in computer vision and multimodal processing under an umbrella of names (abnormal, unusual, or rare events) [17, 19, 6]. In this paper we employ the term *unusual event*, which we define as events with the following properties: (1) they seldom occur (rarity); (2) they may not have been thought of in advance (unexpectedness); and (3) they are relevant for a particular task (relevance).

*This work was supported by the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), and the EC project Augmented Multi-party Interaction (AMI, pub. AMI-62).

It is clear from such a definition that unusual event detection entails a number of challenges. The rarity of an unusual event means that collecting sufficient training data for supervised learning will often be infeasible, necessitating methods for learning from small numbers of examples. In addition, more than one type of unusual event may occur in a given data sequence, where the event types can be expected to differ markedly from one another. This implies that training a single model to capture all unusual events will generally be infeasible, further exacerbating the problem of learning from limited data. As well as such modeling problems due to rarity, the unexpectedness of unusual events means that defining a complete event lexicon will not be possible in general, especially considering the genre- and task-dependent nature of event relevance.

Most existing works on event detection have been designed to work for specific events, with well-defined models and prior expert knowledge, and are therefore ill-posed for handling unusual events. Alternatives to these approaches, addressing some of the issues related to unusual events, have been proposed recently [17, 19, 6]. However, the problem remains unsolved.

In this paper, we propose a framework for unusual event detection. Our approach is motivated by the observation that, while it is unrealistic to obtain a large training data set for unusual events, it is conversely possible to do so for usual events, allowing the creation of a well-estimated model of usual events. In order to overcome the scarcity of training material for unusual events, we propose the use of Bayesian adaptation techniques [14], which adapt a usual event model to produce a number of unusual event models in an unsupervised manner. The proposed framework can thus be considered as a semi-supervised learning technique.

In our framework, a new unusual event model is derived from the usual event model at each step of an iterative process via Bayesian adaptation. Temporal dependencies are modeled using HMMs, which have recently shown good performance for unsupervised learning [1]. We objectively evaluate our algorithm on a number of audio, visual, and audio-visual data streams, each generated by a separate source, and containing different events. With relatively simple audio-visual features, and compared to both supervised and unsupervised baseline systems, our framework produces encouraging results.

The paper is organized as follows. Section 2 describes related work. The proposed framework is introduced in Section 3. In Section 4, we present experimental results and discuss our findings. We conclude the paper in Section 5.

2 Related Work

There is a large amount of work on event detection. Most works have been centered on the detection of predefined events in particular conditions using supervised statistical learning methods, such as HMMs [12, 7, 18], and other graphical models [3, 11, 10, 9]. In particular, some recent work has attempted to recognize *highlights* in videos, e.g., sports [15, 7, 18]. In our view, this concept is related but not identical to unusual event detection. On one hand, typical highlight events in most sports can be well defined from the sports grammar and, although rare, are predictable (e.g., goals in football, home-runs in baseball, etc). On the other hand, truly unusual events (e.g. a blackout in the stadium) could certainly be part of a highlight.

Fully supervised model-based approaches are appropriate if unusual events are well-defined and enough training samples are available. However, such conditions often do not hold for unusual events, which render fully supervised approaches ineffective and unrealistic. To deal with the problem, an HMM approach was proposed in [6] to detect unusual events in aerial videos. Without any models for usual activities, and with only one training sample, unusual events models are handcoded using a set of predefined spatial semantic primitives (e.g. "close" or "adjacent"). Although unusual event models can be created with intuitive primitives for simple cases, it is infeasible for complex events, in which primitives are difficult to define.

As an alternative, unsupervised approaches for unusual event detection have also been proposed [17, 19]. In a far-field surveillance setting, the use of co-occurrence statistics derived from motion-based features was proposed in [17] to create a binary-tree representation of common patterns. Unusual events were then detected by measuring aspects of how usual each observation sequence was. The work in [19] proposed an unsupervised technique to detect unusual human activity events in a surveillance setting, using analysis of co-occurrence between video clips and motion / color features of moving objects, without the need to build models for usual activities.

Our work attempts to combine the complementary advantages of supervised and unsupervised learning in a probabilistic setting. On one hand, we learn a general usual event model exploiting the common availability of training data for such an event type. On the other hand, we use Bayesian adaptation techniques to create models for unusual events in an iterative, data-driven fashion, thus addressing the problem of lack of training samples for unusual events, without relying on pre-defined unusual event sets.

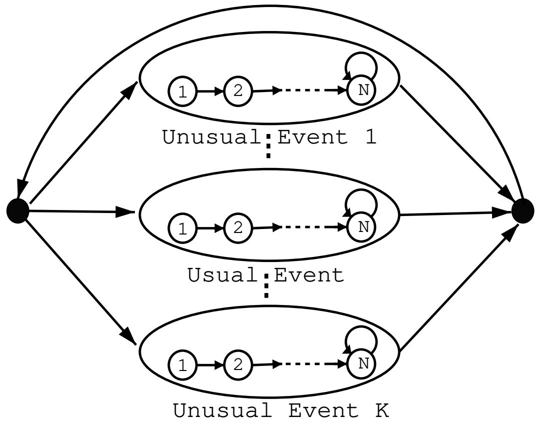

Figure 1. HMM topology for the proposed framework

3 Iterative Adapted HMM

In this section, we first introduce our computational framework. We then describe the implementation details.

3.1 Framework Overview

As shown in Figures 1 and 3, our framework is a hierarchical structure based on an ergodic K-class Hidden Markov Model (HMM) (K is the number of unusual event states plus one usual event state), where each state is a sub-HMM with minimum duration constraint. The central state represents usual events, while the others represent unusual events. All states can reach (or be reached from) other states in one step, and every state can transmit to itself.

Our method starts by having only one state representing usual events (Figure 2, step 0). It is normally easy to collect a large number of training samples for usual events, thus obtaining a well-estimated model for usual events. A set of parameters θ^* of the usual-event HMM model is learned by maximizing the likelihood of observation sequences $\{X_1, X_2, ..., X_M\}$ as follows:

$$\theta^* = \arg\max_{\theta} \prod_{j=1}^{M} P(X_j|\theta). \quad (1)$$

The probability density function of each HMM state is assumed to be a Gaussian Mixture Model (GMM). We use the standard Expectation-Maximization (EM) algorithm [5] to estimate the GMM parameters. In the E-step, a segmentation of the training samples is obtained to maximize the

0. **Training the general model**
A general usual event model is estimated with a large number of training samples.
1. **Outlier detection**
Slice the test sequence into fixed length segments. The segment with the lowest likelihood given the general model is identified as outlier.
2. **Adaptation**
A new unusual event model is adapted from the general usual event model using the detected outlier.
The usual event model is adapted from the general usual event model using the other segments.
3. **Viterbi decoding**
Given a new HMM topology (with one more state), the test sequences are decoded using Viterbi algorithm to determine the boundary of events.
4. **Outlier detection**
Identify a new outlier, which has the smallest likelihood given the adapted usual event model.
5. Repeat step 2, 3, 4
6. **Stop**
Stop the process after the given number of iterations.

Figure 2. Iterative adapted HMM

likelihood of the data, given the parameters of the GMMs. This is followed by an M-step, where the parameters of the GMMs are re-estimated based on this segmentation. This creates a general usual event model.

Given the well-estimated usual event model and an unseen test sequence, we first slice the test sequence into fixed length segments with overlapping. This is done by moving a sliding window. The choice of the sliding window size corresponds to the minimum duration constraint in the HMM framework. Given the usual event model, the likelihood of each segment is then calculated. The segment with the lowest likelihood value is identified as an outlier (Figure 2, step 1). The outlier is expected to represent one specific unusual event and could be used to train an unusual event model. However, one single outlier is obviously insufficient to give a good estimate of the model parameters for unusual events. In order to overcome the lack of training material, we propose the use of model adaptation techniques, such as Maximum a posteriori (MAP) [14], where we adapt the already well-estimated usual event model to a particular unusual event model using the detected outlier, i.e, we start from the usual event model, and move towards an unusual event model in some constrained way (see Section 3.2 for implementation details). The original usual event model is trained using a large number of samples, which generally means that it yields Gaussians with relatively large variances. In order to make the model better suited for test se-

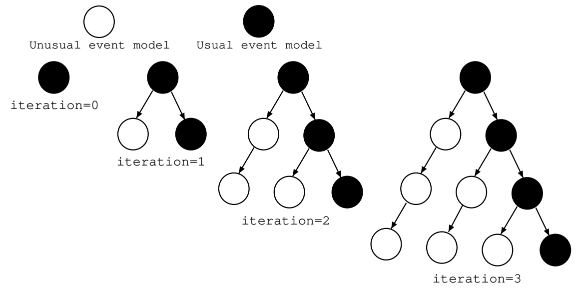

Figure 3. Illustration of the algorithm flow. At each iteration, two leaf nodes, one representing usual events and the other one representing unusual events, are split from the parent usual event node; A leaf node representing an unusual event is also adapted from the parent unusual event node.

quences, the original usual event model is also adapted with the other segments (except for the detected outlier), using the same adaptation technique for the unusual event model (Figure 2, step 2).

Given the new unusual and usual event models, both adapted from the general usual event model, the HMM topology is changed with one more state. Hence the current HMM has 2 states, one representing the usual events and one representing the first detected unusual event. The Viterbi algorithm is then used to find the best possible state sequence which could have emitted the observation sequence, according to the maximum likelihood (ML) criterion (Figure 2, step 3). Transition points, which define new segments, are detected using the current HMM topology and parameters. A new outlier is now identified by sorting the likelihood of all segments given the usual event model (Figure 2, step 4). The detected outlier provides material for building another unusual event model, which is also adapted from usual event model. At the same time, both the unusual and usual event models are adapted using the detected unusual / usual event samples respectively. The process repeats until we obtain the desired number of unusual events. At each iteration, all usual / unusual event models are adapted from the parent node (see Figure 3), and a new unusual event model is derived from the usual event model via Bayesian adaptation. The number of iterations thus corresponds to the number of unusual event models, as well as the number of states in the HMM topology.

As shown in Figure 3, the proposed framework has a top-down hierarchical structure. Initially, there is only one node in the tree, representing the usual event model. At the first iteration, two new leaf nodes are split from the upper parent node: one representing usual events and the other one representing unusual events. At the second iteration, there are three leaf nodes in the tree: two for unusual events and one for usual events. The tree grows in a top-down fashion until we reach the desired number of iterations. The proposed algorithm is summarized in Figure 2.

Compared with previous work on unusual event detection, our framework has a number of advantages. Most existing techniques using supervised learning for event detection require manually labeling of a large number of training samples. As our approach is semi-unsupervised, it does not need explicitly labeled unusual event data, facilitating initial training of the system and hence application to new conditions. Furthermore, we derive both unusual event and usual event models from a general usual event model via adaptation techniques in an online manner, thus allowing for a faster model training. In addition, the minimum duration constraint for temporal events can be easily imposed in the HMM framework by simply changing the number of cascaded states within each class.

In the next subsection, we give more details on the used adaptation techniques.

3.2 MAP Adaptation

Several adaptation techniques have been proposed for GMM-based HMMs, such as Gaussian clustering, Maximum Likelihood Linear Regression (MLLR) and Maximum a posteriori (MAP) adaptation (also known as Bayesian adaptation) [14]. These techniques have been widely used in tasks such as speaker and face verification [14, 4]. In these cases, a general world model of speakers / faces are trained and then adapted to the particular speaker / face. In our case, we train a general usual event model and then use MAP to adapt both unusual and usual event models.

According to the MAP principle, we select parameters θ^* such that they maximize the posterior probability density, that is:

$$\begin{aligned} \theta^* &= \arg\max_\theta P(\theta|X) \\ &= \arg\max_\theta P(X|\theta) \cdot P(\theta), \end{aligned} \quad (2)$$

where $P(X|\theta)$ is the data likelihood and $P(\theta)$ is the prior distribution. When using MAP adaptation, different parameters can be chosen to be adapted [14]. In [14, 4], the parameters that are adapted are the Gaussian means, while the mixture weights and standard deviations are kept fixed and equal to their corresponding value in the world model. In our case we adapt all the parameters. The reason to adapt the weights is that we model events (either usual or unusual) with different components in the mixture model. When only one specific event is present, it is expected that the weights of the other components will be adapted to zero (or a relatively small value). We also adapt the variances in order to move from the general model, which may have larger covariance matrix, to a specific model, with smaller variance, focusing on one particular event in the test sequence.

Following [14], there are two steps in adaptation. First, estimates of the statistics of the training data are computed for each component of the old model. We use $\{w_i^{new}, \mu_i^{new}, \sigma_i^{new}\}$ to represent the weight, mean and variance for component i in the new model, respectively. These parameters are estimated by ML, using the well-known equations [2],

$$w_i^{new} = \frac{1}{M}\sum_{j=1}^{M} P(i|x_j, \theta), \quad (3)$$

$$\mu_i^{new} = \frac{\sum_{j=1}^{M} x_j P(i|x_j, \theta)}{\sum_{j=1}^{M} P(i|x_j, \theta)}, \quad (4)$$

$$\sigma_i^{new} = \frac{\sum_{j=1}^{M} P(i|x_j, \theta)(x_j - \mu_i^{new})(x_j - \mu_i^{new})^T}{\sum_{j=1}^{M} P(i|x_j, \theta)}, \quad (5)$$

where M is the number of data examples.

In the second step, the parameters of a mixture i are adapted using the following set of update equations [8].

$$\hat{w}_i = \alpha \cdot w_i^{old} + (1-\alpha) \cdot w_i^{new}, \quad (6)$$

$$\hat{\mu}_i = \alpha \cdot \mu_i^{old} + (1-\alpha) \cdot \mu_i^{new}, \quad (7)$$

$$\begin{aligned} \hat{\sigma}_i = \alpha \cdot (\sigma_i^{old} + (\hat{\mu}_i - \mu_i^{old})(\hat{\mu}_i - \mu_i^{old})^T) \\ +(1-\alpha) \cdot (\sigma_i^{new} + (\hat{\mu}_i - \mu_i^{new})(\hat{\mu}_i - \mu_i^{new})^T), \end{aligned} \quad (8)$$

where $\{\hat{w}_i, \hat{\mu}_i, \hat{\sigma}_i\}$ are weight, mean and variance of the adapted model in component i, $\{w_i^{old}, \mu_i^{old}, \sigma_i^{old}\}$ are the corresponding parameters in the old component i respectively, and α is a weighting factor to control the balance between old model and new estimates. The smaller the value of α, the more contribution the new data makes to the adapted model.

4 Experiments and Results

In this section, we first introduce the performance measures and baseline systems we used to evaluate our results. Then we illustrate the effectiveness of the proposed framework using audio, visual and audio-visual events.

4.1 Performance Measures

The problem of unusual event detection is a two-class classification problem (unusual events *vs.* usual events), with two types of errors: a *false alarm* (FA), when the method accepts an usual event sample (frame), and a *false rejection* (FR), when the method rejects an unusual event sample. The performance of the unusual event detection method can be measured in terms of two error rates: the *false alarm rate* (FAR), and the *false rejection rate* (FRR), defined as follows:

$$\text{FAR} = \frac{\text{number of FAs}}{\text{number of usual event samples}} \times 100\%, \quad (9)$$

$$\text{FRR} = \frac{\text{number of FRs}}{\text{number of unusual event samples}} \times 100\%. \quad (10)$$

The performance for an ideal event detection algorithm should have low values of both FAR and FRR. We also use the *half-total error rate* (HTER), which combines FAR and FRR into a single measure: HTER $= \frac{\text{FAR}+\text{FRR}}{2}$.

4.2 Baseline Systems

To evaluate the results, we compare the proposed semi-supervised framework with the following baseline systems.

Supervised HMM: Two standard HMM models, one for usual events and one for unusual events, are trained using manually labeled training data according to Equation 1. For testing, the event boundary is obtained by applying Viterbi decoding on the sequences.

For supervised HMM, we test two cases. In the first case, we train usual and unusual event models using a large (sufficient) number of samples, referred to as *supervised-1*. In the second case, referred to as *supervised-2*, around 10% of the unusual event training samples from the first case are used to train the unusual event HMM. The purpose of *supervised-2* is to investigate the case where there is only a small number of unusual event training samples.

Unsupervised HMM: The second baseline system is an agglomerative HMM-based clustering algorithm, recently proposed for speaker clustering [1], and that has shown good performance. The unsupervised HMM clustering algorithm starts by over-clustering, i.e. clustering the data into a large number of clusters. Then it searches for the best candidate pair of clusters for merging based on the criterion described in [1]. The merging process is iterated until there are only two clusters left, one assumed to correspond to usual events, and another one for unusual events. We assume that the cluster with the largest number of samples represents usual events, and the other cluster represents unusual events. This model is referred to as *unsupervised*.

For both the proposed approach and the baseline methods, all parameters are selected to minimize *half-total error rate* (HTER) criterion on a validation data set.

4.3 Results on Audio Events

For the first experiment, we used a data set of audio events obtained through a sound search engine [1]. The purpose of this experiment is to have a controlled setup for evaluation of our algorithm. We first selected 60 minutes audio data containing only 'speaking' events. We then manually mixed it with other interesting audio events, namely 'applause', 'cheer', and 'laugh' events. The length of each concatenated segment is random. 'Speaking' is labeled as usual event, while all the other events are considered unusual. The minimum duration for audio events is two seconds.

[1] http://www.findsounds.com/types.html

Table 1. Audio events data. Number of frames for various methods (*NA:* Not Applicable).

method	train set		test set	
	usual	unusual	usual	unusual
our approach	90000	*NA*	72750	2250
supervised-1	90000	20000		
supervised-2	90000	2000		
unsupervised	*NA*	*NA*		

We extracted Mel-Frequency Cepstral Coefficients (MFCCs) features for this task. MFCC are short-term spectral-based features and have been widely used in speech recognition [13] and audio event classification. We extracted 12 MFCC coefficients from the original audio signal using a sliding window of 40ms at fixed intervals of 20ms. The number of training and testing frames for the different methods is shown in Table 1. Note that there is no need for unusual event training data for our approach. For the unsupervised HMM, there is no need for training data. The percentage of frames for unusual events in the test sequence is around 3%.

Figure 4(a) shows the performance of the proposed approach with respect to the number of iterations. We observe that FRR always decreases while FAR continually increases with the increase of the number of iterations. This is because our approach derives a new unusual event modal from the usual event model via Bayesian adaptation at each iteration. With the increase of unusual event models, more unusual events can be detected, while more usual events were falsely accepted as unusual events.

Figure 4(b) shows the performance comparison between the proposed approach and baseline systems in terms of HTER. We can see that the supervised HMM with sufficient amount of training data gives the best performance. The proposed approach improves the performance, compared to the *supervised-2* and *unsupervised* baselines. The results show that the benefit of using the proposed approach is not performance improvement when sufficient training data is available, but rather its effectiveness when there are not enough training samples for unusual events. The best result of our approach is obtained at 4 iterations (HTER = 6.65%), slightly worse than *supervise-1* (HTER = 5.29%), showing the effectiveness of our approach given that it does not need any unusual event training data.

4.4 Results on Visual Events

The visual data we investigate is a 30-minute long poker game video, containing 26 different events and originally manually labeled and used in [19]. Seven cheating related events, including 'hiding a card', 'exchanging cards', 'passing cards under table', etc., are categorized as unusual

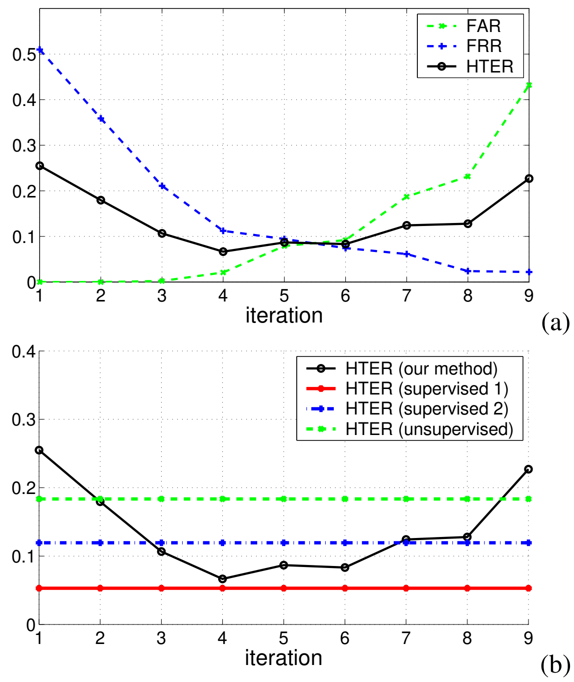

Figure 4. Results for audio unusual event detection. The X-axis represents the number of iterations in our approach.

events (see Figure 6). Other events such as 'playing cards', 'drinking water', and 'scratching', are considered as usual events. The minimum duration for these visual events is 15 frames.

The number of training and testing frames for different methods is shown in Table 2. While we chose this visual task to show application on an existing data set, we note that the percentage of frames of unusual events in the test sequence is about 17%, which does not correspond very well to the assumption of *rarity* made by our model. The unusual event testing data for the *supervised-1* method is much smaller, compared with other methods. This is because we use a larger number of unusual event frames (1320) for training, and we are left with a small number of unusual event frames (195) for testing. To deal with this problem, we repeat experiments for *supervised-1* ten times by randomly splitting total unusual events into two parts: one with 1320 frames for training, and the other one with 195 frames for testing. We report the mean results of the ten runs. Note also that the amount of training data for the unusual model (1320 frames) is smaller than the previous experiments.

We extract motion and color features from moving blocks of each frame in the video in a similar way as in [19]. We start with a static background image. We detect the moving objects using background substraction. We then superimpose a 6×6 grid on the detected motion mask. We first compute a motion histogram. In each tile of the grid, we calculate the total number of motion pixels, and

Table 2. Video events data. Number of frames for various methods (*NA:* Not Applicable).

method	train set		test set	
	usual	unusual	usual	unusual
our approach	9000	*NA*	7387	1515
supervised-1	9000	1320		195
supervised-2	9000	300		1215
unsupervised	*NA*	*NA*		1515

these features are concatenated to form a $6 \times 6 = 36$ dimension feature vector to describe the motion in the current frame. In a similar way, we can compute the color histogram for the moving objects in chromatic color space (defined by $r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B}$). We concatenate the motion histogram and the color histogram into a $108 = 36 + 2 \times 36$ dimension feature vector. To reduce the feature space dimension and for feature decorrelation, we apply a Principal Component Analysis (PCA) to transform the 108-dimensional features to 36-dimensional features.

The results are shown in Figure 5. Overall, this is a more difficult task. We observe the similar trend of FAR and FRR as in audio event detection, with respect to the number of iterations in our approach. The best result of our approach is obtained with 4 iterations, although the values of HTER are relatively stable between 4 iterations and 7 iterations. We come to similar conclusions as for the audio event detection, that is, the supervised approach with sufficient training samples provides the best performance, while the proposed framework is better than the other baseline systems. Note that the supervised approach with small number of training samples performs worse than the unsupervised approach.

4.5 Results on Audio-Visual Events

We also apply our framework to audio-visual unusual event detection using the ICCV'03 recorded presentation videos, publicly available [2]. Each presentation video is about 20 minutes in length with 25 frames per second. We define a set of multimodal unusual events, including 'speaker showing demo, audience applause', 'speaker playing video, audience laugh', and 'speaker interrupted by audience's questions'. Note that since some unusual events in the presentation setting cannot be defined before watching the entire database, the unusual events list we define here should be regarded as a small subset.

A set of audio-visual features were extracted. For audio features, we use the same features as in section 4.3. For visual features, we extract a motion histogram from each frame of the video, computed in a similar way to section 4.4. Audio and visual features were then concatenated.

Since the occurrence of unusual events is rare, manually labeling a large amount of samples is impractical, high-

[2] http://www.robots.ox.ac.uk/~awf/iccv03videos

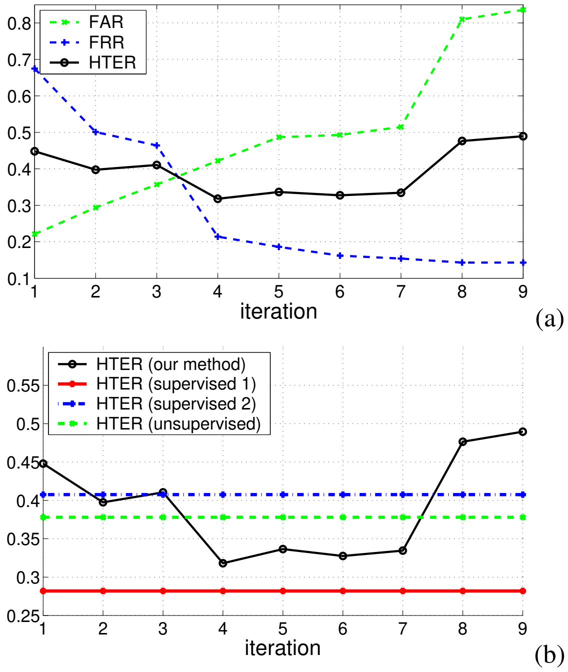

Figure 5. Results of visual unusual events detection.

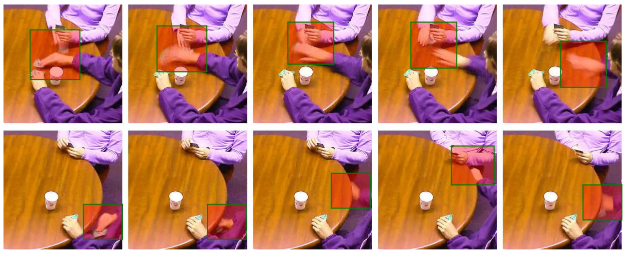

Figure 6. Top: Visual event of ‘exchanging cards’; Bottom: Visual event of ‘passing cards under table’

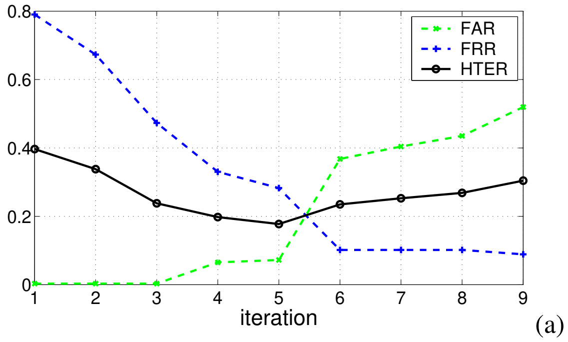

Figure 7. Results of our approach in terms of FAR, FRR and HTER.

Table 3. Overall the best results

Events	Method	FAR %	FRR %	HTER %
audio	our method	2.09	11.2	6.65
	supervised 1	3.97	6.62	5.29
	supervised-2	11.8	12.6	12.2
	unsupervised	12.5	24.2	18.3
visual	our method	42.2	21.4	31.8
	supervised-1	26.8	29.6	28.2
	supervised-2	41.3	40.2	40.7
	unsupervised	40.1	35.5	37.8
audio-visual	our approach	7.20	28.2	17.7

lighting the need for semi-supervised or unsupervised approaches. Due to the lack of sufficient annotated training data for the supervised baselines, we only report results of our approach. Two presentation videos are used for training to build the general usual event model. We then apply our framework to a third meeting for unusual event detection. One of the co-authors labeled the events by hand to obtain a ground truth in the three videos. The results are shown in Figure 7. We observe that, with the increase of iterations, FRR decreases while FAR increases, which means that more unusual events are detected, but at the cost of falsely accepting more usual events as unusual events. The best result of our approach is obtained when the number of iterations is 5.

4.6 Overall Discussion

Table 3 summarizes overall results of audio, visual and audio-visual unusual event detection. For the proposed approach, the results correspond to the iteration with the minimum HTER. For both audio and visual unusual event detection, we can see that supervised HMM well-trained with sufficient data achieves the best performance while the proposed approach performs better than the other baseline systems.

As a well-known rule-of-thumb, the number of training samples needed for a well-trained model is directly related with the model complexity (the number of model parameters). The penalty for training with insufficient data is over-fitting, *i.e.* poor generalization capability. Both our approach and the baseline methods are based on HMMs for usual and unusual events modeling and hence have similar model complexity.

For the proposed approach, we currently do not determine the optimal number of iterations. As shown in Figures 4, 5 and 7, finding the optimal number of iterations is a trade-off between FAR and FRR. Some applications require more unusual events detected thus need more iterations. Otherwise, we might stop iterations at the early stages if fewer false alarms are expected. Automatic model selection is a difficult problem that we are studying, in particular with the Bayesian Information Criterion (BIC) [16]. In our approach, there is one additional state in the HMM topology at each iteration, which results in an increase of both the number of model parameters and the likelihood of a test sequence. BIC could be used to handle the trade-off between model complexity and data likelihood.

We also note that feature selection is a critical issue in unusual event detection, particularly when using a semi- or unsupervised approach. The nature of the events found by the system will necessarily relate to the nature of discrimination provided by the features. In the above experiments, while the audio features seem to allow such discrimination, ongoing research should include investigation of different visual features.

Finally, regarding the three properties we used to define an unusual event (rarity, unexpectedness, and relevance), our method aims at accounting for the first two (one could argue that unexpectedness is a feature of some rare events). Relevance is a task-dependent property, whose incorporation in our work would require human intervention.

5 Conclusion

In this paper, we presented a semi-supervised adapted HMM framework for unusual event detection. The proposed framework is well suited for cases in which collecting sufficient unusual event training data is impractical and unusual events cannot be defined in advance. With relatively simple audio-visual features, and compared to both supervised and unsupervised baseline systems, our framework produces encouraging results. In future work, we will investigate the use of some criterion for optimizing the number of iterations, as well as improved feature selection.

Acknowledgments

We thank Hua Zhong (Carnegie Mellon University), Jianbo Shi and Mirko Visontai (University of Pennsylvania) for providing visual data for experiments. We also thank David Barber (IDIAP Research Institute) for helpful comments.

References

[1] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[2] J. Bilmes. A gentle tutorial of the EM algirthm and its application to parameter estimation for gaussian mixture and hidden markov models. ICSI-TR-97-021 U.C. Berkeley, 1997.

[3] H. Buxton and S. Gong. Advanced Visual Surveillance using Bayesian Networks. In *Prof. IEEE ICCV*, 1995.

[4] F. Cardinaux, C. Sanderson, and S. Bengio. Adapted generative models for face verification. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39(B), pp. 1–38*, 1977.

[6] M.T. Chan, A. Hoogs, J. Schmiederer, and M. Perterson. Detecting rare events in video using semantic primitives with HMM. *In Proc. ICPR*, August 2004.

[7] P Chang, M Han, and Y Gong. Highlight detection and classification of baseball game video with hidden markov models. In *Proc. IEEE ICIP*, New York, Sept. 2002.

[8] J. L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains. In *IEEE Transactions on Speech Audio Processing*, volume 2, pp. 291–298, April 1994.

[9] S. Gong and T. Xiang. Recognition of group activities using a dynamic probabilistic network. In *Proc. IEEE ICCV*, Nice, Oct. 2003.

[10] S. Hongeng, F. Bremond, and R. Nevatia. Bayesian framework for video surveillance application. In *Proc. ICPR*, 2000.

[11] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, archive Vol.23(8) August 2001.

[12] N. Oliver, B. Rosario and A. Pentland. A Bayesian Computer Vision System for Modeling Human Interactions. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, archive Vol.22(8) August 2000.

[13] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing, vol. 10, pp. 19–41*, 2000.

[15] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *Proc. ACM Multimedia, pp. 105–115*, Oct. 2000.

[16] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics, vol. 6, pp. 461–464,* 1978.

[17] C. Stauffer, W. Eric, and L. Grimson. Learning patterns of activity using real-time tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, archive Vol.22(8) August 2000.

[18] J Wang, C Xu, E.S. Chng, and Q Tian. Sports highlight detection from keyword sequences using hmm. In *Proc. IEEE ICME*, Taiwan, June 2004.

[19] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. IEEE CVPR*, June. 2004.