Modelling Crowd Scenes for Event Detection

Ernesto L. Andrade[1], Scott Blunsden[2] and Robert B. Fisher[1]
IPAB, School of Informatics, University of Edinburgh
King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, UK
[1]eaneto,rbf@inf.ed.ac.uk, [2]S.J.Blunsden@sms.ed.ac.uk

Abstract

This work presents an automatic technique for detection of abnormal events in crowds. Crowd behaviour is difficult to predict and might not be easily semantically translated. Moreover it is difficulty to track individuals in the crowd using state of the art tracking algorithms. Therefore we characterise crowd behaviour by observing the crowd optical flow and use unsupervised feature extraction to encode normal crowd behaviour. The unsupervised feature extraction applies spectral clustering to find the optimal number of models to represent normal motion patterns. The motion models are HMMs to cope with the variable number of motion samples that might be present in each observation window. The results on simulated crowds demonstrate the effectiveness of the approach for detecting crowd emergency scenarios.

1 Introduction

In recent years computer vision and machine learning techniques have been applied to modeling and recognition of human activities and interactions. The application domains for these techniques usually involve simple environments such as offices [8], kitchens [3] , cargo bays [6] and loading docks [5] such that activity recognition is focused upon modeling the actions and interactions of small groups of people/objects. However, there have been few attempts to model larger groups of people, crowds, which are mostly based on discriminative classifiers [10]. The analysis of crowd movements and behaviour is of particular interest in the surveillance domain [7]. In scenarios where hundreds of cameras are monitored by a few operators behavioural analysis of crowds is useful as a tool for video pre-screening.

In order to model a crowd the model must cope with a large variation in densities and motions present in a real crowd. This requires a huge amount of data to enable good supervised/unsupervised learning for discriminative or generative crowd models. Moreover in the surveillance domain usually there are no examples of the emergency/abnormal events to be analysed. Thus the first assumption for our crowd modelling is that we are trying to model the degree of similarity between the trained model and the new unseen video data. Therefore the events are classified as normal or abnormal behaviour without having any other particular labels for them. This arrives from the fact that crowds are difficult to treat semantically. In a real crowd scene one can not beforehand easily specify or train particular labels for behavioural analysis. This would discretise the input space and thus simplify the analysis. However, unsupervised learning techniques provide the means to learn the typical labels (space-time behavioural patterns) and have been applied for similar problems in video analysis [15] [12]. In our work we apply projections on the principal components of the training flow fields as features for the learning algorithms. The automatic feature extraction involves fitting an HMM for each video segment and performing spectral clustering using the similarity matrix computed using inter-segment likelihoods. The resulting clustered video segments are used to train a new set of HMMs which representing the optical flow variations on the normal example set. Abnormality detection is based on a threshold on the HMM bank likelihood function. This framework is applied to detect simulated emergencies in crowds.

2 Related Work

The use of principal component analysis of optical flow fields as features is demonstrated in [4], where principal components of video sequences are used to construct a linear basis for complex motion phenomena. Unusual events are analysed in a similar context in [6] and [14] where deviations from example normal behaviour are used to characterise abnormality. Spectral clustering using HMMs as similarity measures is used for trajectory classification in [9]. In another related spectral clustering application it is used to automatically determine models for video sequence in [12]. Our approach is based on the general concepts in these references and to the best of our knowledge this work is the first combined application of these techniques to the problem of abnormality detection in crowds.

3 Overview of Modeling Events in Crowds

The characterisation of normal behaviour for the crowd uses the normal optical flow patterns to estimate the model parameters. The modelling process involves four phases: 1) Preprocessing: background modelling and optical flow computation; 2) Feature prototypes: principal components analysis on the example flow fields, 3) Spectral Clustering: automatic determination of the number of HMMs to represent the flow sequences and 4) Bank of models: training of the HMM models using the data of each cluster per model. The analysis concentrates on identifying unusual events in the crowd by comparing the new observation's likelihood to a detection threshold. Details of this are given in the next section.

4 Unsupervised Extraction of Video Prototypes

4.1 Preprocessing

Preprocessing involves the construction of a Mixture of Gaussians background model for the scene based on [11]. The background model produces a mask with the detected foreground objects per frame. In parallel to foreground extraction robust optical flow is computed for the whole frame using the techniques described in [2]. Prior to the optical flow computation the sequence is smoothed with a 5x5x5 Gaussian filter to reduce acquisition noise ($\sigma = 0.8$). The resulting optical flow is sub-sampled by a median filter with a window of size 8x8 applied independently to the horizontal and vertical components. The combination of flow information with the foreground mask allows the analysis to only consider flow vectors inside foreground objects, reducing observation noise. The motion parameters are encoded in a sample vector of the form $\mathbf{s} = (u, v)$, where u and v are horizontal and vertical optical flow components. Prior to the analysis the foreground mask is superimposed to the optical flow output resulting in the motion parameters for the detected foreground objects. All values outside the foreground mask are set to zero to characterise the static regions.

4.2 Video Segmentation

The assumption for video segmentation is that there is no distinctive activity or periods of inactivity everywhere in the training crowd video. Therefore all segments in the video stream are equally important for prototype extraction. The video sequence $\mathbf{V}$ is segmented in N video segments $\mathbf{V} = \{\mathbf{v_1}, ..., \mathbf{v_N}\}$ of equal length T, $\mathbf{v_n} = \{\mathbf{F_{n1}}, ..., \mathbf{F_{nT}}\}$ as in [15]. $\mathbf{F_{nt}} = (s_1, ..., s_P)$, where P is the number of flow vectors in each frame. $T = 100$ frames (4 seconds) is assumed in the experimental section to contain enough crowd movement for comparison.

4.3 Feature Prototypes

The first step of the prototype extraction is to perform principal component analysis (PCA) on the optical flow fields of each frame $\mathbf{F_{nt}} = ((u_1, v_1), ..., (u_P, v_P))$ in $\mathbf{V}$. The first J eigenvectors with the largest eigenvalues are selected to form a basis for the projection. The projection reduces the input feature dimensionality from the dimension of flow fields samples P to the dimension of the selected eigenvectors J. The resulting set of feature vectors for the n$-th$ segment in $\mathbf{V}$ is:

$$\mathbf{W_n} = \{\mathbf{w_{n1}}, ..., \mathbf{w_{nT}}\} \tag{1}$$

where $\mathbf{w_{nt}}$ is a vector representing the projection of the t-th frame in the n-th segment over the selected eigenvectors, defined as

$$\mathbf{w_{nt}} = \{g_{nt1}, ..., g_{ntm}\} \;\; m \;=\; 1 \ldots J \tag{2}$$

where g_{ntm} is the weight associated with the m-th eigenvector. The vectors in (1) represent the activity pattern in the n-th segment.

4.4 Spectral Clustering

The derivation of the similarity measure of the video segments for spectral clustering is based on likelihood of the observations inside the segments given by a Hidden Markov model. For that a Multiple Observation Hidden Markov Model (MOHMM) [12] is trained with the feature vectors in each video segment inside the training set resulting in $\mathbf{B_k}, k = 1..N$ models. This MOHMM structure is ergodic with J states (same as the number of selected eigenvectors) and one Gaussian emission probability per state in order to reduce the number of samples needed to train the MOHMM (assuming independence in the input space of eigenvectors projections).

The measure of similarity between video segments is defined as:

$$S_{ij} = \frac{1}{2}\{log\ P(\mathbf{W_j}|\mathbf{B_i}) + log\ P(\mathbf{W_i}|\mathbf{B_j})\} \tag{3}$$

The pairwise similarity values between the video segments form a similarity matrix $\mathbf{S}$. The similarity matrix is subject to spectral clustering using the algorithm described [13] to automatically find the number of groups in the video data.

4.5 HMM Training

After spectral clustering the video segments are grouped in K different classes. All the samples $\mathbf{W_n}$ in each class are used to train a new MOHMM per class $\mathbf{M}_k$. The final model for the video sequence has the form:

$$P(\mathbf{W}|\mathbf{M}) = \sum_{k=1}^{K} \frac{N_k}{N} P(\mathbf{W}|\mathbf{M}_k) \tag{4}$$

where N_k is the number of video segments clustered in the class k and the ratio represents a prior on the model weights.

4.6 Event Classification

The classification of normal and abnormal events is based on the comparison of the current observation's likelihood given by the bank of MOHMM models and the detection threshold. The observation of the n-th test video segment $\mathbf{W^o_k}$ (eg. the previous 50 frames) is considered abnormal if:

$$P(\mathbf{W^o_k}|\mathbf{M}) < Th_{Ab} \tag{5}$$

The test video features are extracted by projecting the test flow fields on the J eigenvectors of the sub-space derived from the training set.

5 Experimental Results

5.1 Crowd Simulation Data

There are two simulated data sets: normal flow and blocked exit. In the normal flow simulation a crowd flows in one direction in the scene. In the blocked exit simulation the crowd cannot leave the scene and starts to press each against the exit. The simulation technique is described in [1]. The original frame size is 384x288 pixels and the optical flow observations are decimated, by the u and v median over 8x8 blocks, resulting in optical flow image of 48 x 36 (P=1728) flow vectors. Using $J = 10$ eigenvectors gives a total input space of 10 elements per frame. One of the eigenvectors of the optical flow fields used for feature extraction is shown in Fig.1. The normal simulated sequence has 5000 frames and is divided for clustering in $N = 50$ segments of size $T = 100$. The similarity matrix for the training set is shown in Fig.2. It displays a high degree of inter-segment similarity on the crowd video, which is due mainly to the high density of the simulated crowd. $K = 10$ video segment clusters are automatically selected by the spectral clustering algorithm. The results for the self-likelihood of the training sequence in the trained model bank are shown in Fig.3.(a). The results for the detection of the blocked exit emergency event are shown in Fig.3.(b). There is a clear and quick drop in the likelihood function less than 100 frames (4 seconds) after the exit blocking. The size of the observation window used to compute the likelihood in Fig.3 is 50 frames. Larger window sizes tend to smooth the likelihood function reducing the sensitivity of the detector. The detection threshold Th_{Ab} is defined as a value smaller than the minimum likelihood value present in the normal training set. In order to evaluate the influence of the number of eigenvectors J in producing a likelihood drop to be detected using Th_{Ab} another experiment using 5 independent simulation runs for the blocked exit event are produced, with 3000 frames per sequence and the blocking occurring at frame 2000. The comparison is based on the average likelihood for each sequence before and after the event and is summarised in Table 1. The trend shows that the best detection performance is achieved with $J = 10$. However, it is necessary to correctly select the appropriate number of eigenvectors carefully.

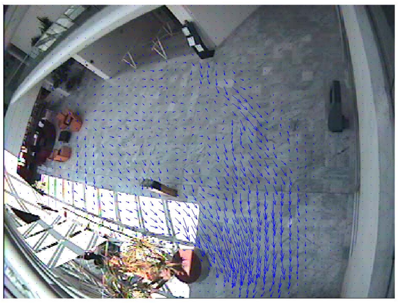

Figure 1. Eigenflows for the normal training set (first eigenvector).

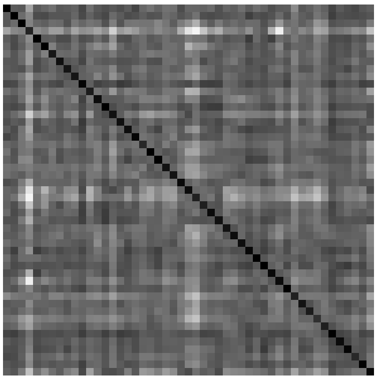

Figure 2. Similarity matrix for the video segments in the training set S_{ij}. Darker blocks indicate higher similarity.

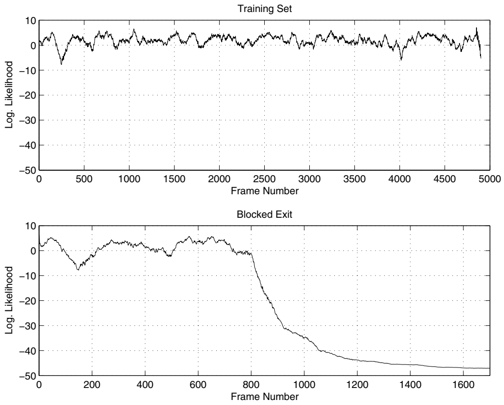

Figure 3. Log-likelihood results for: (a) the training set and (b) blocked exit event (at frame 800, J = 10 eigenvectors).

J(# eigv.)	2	4	6	8	10
Before	-1.4129	0.2122	1.9294	3.1639	2.5341
After	-34.7176	-33.5367	-23.6015	-33.3152	-49.5524

Table 1. Blocked exit. Loglikelihood mean before and after the event as a function of the number of eigenvectors J.

6 Conclusion

This work presented an automatic technique for detection of abnormal events in crowds. Using projections of the eigenvectors in a sub-space spanned by the normal crowd scene as an input feature the proposed technique applies spectral clustering to automatically identify the number of distinct motion segments in the sequence. The features in the clustered motion segments are used to train different MOHMMs for the normal sequence, which compose a bank of models for the training simulated video. The experiments show that the bank of models is effective in quickly detecting the simulated emergency situtation in a dense crowd. The investigation of the relation between the number of eigenvectors and the model likelihood variations indicates that all configurations present a significant drop relative to the normal case and are able to correctly detect the emergencies.

Acknowledgements

This work is funded by EPSRC's BEHAVE project GR/S98146.

References

[1] E. L. Andrade and R. B. Fisher. Simulation of crowd problems for computer vision. *First International Workshop on Crowd Simulation*, (3):71–80, 2005.

[2] M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. *4^{th} International Conference on Computer Vision*, pages 231–236, 1993.

[3] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 1:838–845, 2005.

[4] D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson. Design and use of linear models for image motion analysis. *International Journal of Computer Vision*, 36(3):171–193, 2000.

[5] S. Gong and T. Xiang. Recognition of group activities using a dynamic probabilistic network. *Proceedings of the IEEE International Conference on Computer Vision*, pages 742–749, 2003.

[6] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and explanation of anomalous activities: representing activities as bags of event n-grams. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 1:1031–1038, 2005.

[7] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviours. *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, 43:334–352, 2004.

[8] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96:163–180, 2004.

[9] F. Porikli. Learning object trajectory patterns by spectral clustering. *Proceedings IEEE International Conference on Multimedia and EXPO (ICME)*, pages 1171–1174, 2004.

[10] A. L. S. A. Velastin, B. A. Boghossian. Detection of potentially dangerous situations involving crowds using image processing. *Proceedings of the Third ICSC Symposia on Intelligent Industrial Automation (IIA'99) and Soft Computing*, 1999.

[11] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[12] T. Xiang and S. Gong. Video behaviour profiling and abnormality detection without manual labelling. *Proceedings IEEE International Conference on Computer Vision*, 2005.

[13] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA*, 2005.

[14] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted hmms for unusual event detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 1:611–618, 2005.

[15] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, 2:819–826, 2004.