# Abnormal Activity Recognition Using Spatio-Temporal Features

K. G. Manosha Chathuramali, Sameera Ramasinghe, Ranga Rodrigo
Department of Electronic and Telecommunication Engineering
University of Moratuwa
Sri Lanka
Email: manosha@ent.mrt.ac.lk, samramasinghe@gmail.com, ranga@uom.lk

***Abstract*—Abnormal activity detection plays an important role in many areas such as surveillance, military installations, and sports. Existing abnormal activity detectors mostly rely on motion data obtained over a number of frames to characterize abnormality. However, only motion may not be able to capture all forms of abnormality, in particular, poses that do not amount to motion "outliers". In this paper, we propose two different spatio-temporal descriptors, a silhouette and optic flow based method and a dense trajectory based method which additionally include trajectory shape descriptor, to detect abnormalities. These two descriptors enable us to classify abnormal versus non-abnormal activities using SVM. Comparison with existing methods, using five standard datasets, shows that dense trajectory based method outperforms state-of-the-art results in crowd dataset and silhouette and optic flow based method outperforms others in some datasets.**

***Keywords*—Abnormal activity detection, dense trajectories, HOG, HOF, MBH, SVM**

## I. Introduction

Detecting actions is a challenging and interesting area in computer vision. Among them, abnormal activity detection plays an important role in intelligent visual surveillance and smart camera systems. Effective monitoring of public places, such as, railway stations, shopping malls, crowded sports arenas, and military installations needs intelligent visual surveillance. This is due to, for example, increased global security concerns or for use in smart healthcare facilities for daily activity monitoring and elderly monitoring.

In the context of visual surveillance, the definition of the word "abnormal" is "deviating from the ordinary type". According to the definition, abnormal events can be identified as irregular events in comparison to normal ones. However, choosing visual features or descriptors that discriminate abnormal events from regular events remains a challenge. In this paper, we address the above question in the context of recognizing activities in different types of datasets and catalog abnormalities in them. In particular, our task is to identify frames that amount to abnormal events in a given video frame sequence. Although several methods have been proposed to identify abnormalities recently, it remains a challenging problem [1], [2]. Abnormalities can even be hard to define. They entail various types of challenges: The foremost challenge is that "unusual" or "abnormal" things normally occur with unpredictable variations, making it hard to discriminate a truly abnormal event from noisy normal observations. Furthermore the rarity of an abnormal event means that collecting sufficient training data for supervised learning is quite hard.

One traditional approach of abnormal event detection is based on trajectory modeling [3], [4]. It includes tracking each object in the scene, and learning models for the resulting object tracks. Both operations are quite hard when targets are many. The most popular technique that circumvents this problem is dense optic flow, or some other form of spatio-temporal gradients [5], [2], [6]. Adam *et al.* [5] describe probabilities of optical flow in local regions using histograms. Mehran *et al.* [6] draw inspiration from classical studies of crowd behavior that characterize it, using concepts such as social force. Kim and Grauman [2] represent local optical flow patterns with a mixture of probabilistic PCA models, and enforce global consistency using a Markov Random Field (MRF). All these concepts work on optic flow measures of interaction within crowds, which are combined with a Latent Dirichlet Allocation (LDA), as unsupervised learning model for abnormal event detection. Besides that, Zhang *et al.* [7] propose a semi-supervised adapted Hidden Markov Model (HMM) framework, in which normal event models are first learned from a large amount of training data, and unusual event models are learned by a Bayesian adaption in an unsupervised manner. Yong *et al.* [8] propose a method via a sparse reconstruction over the local spatio-temporal patches to detect both local and global abnormal events.

Overall, there is a great diversity of approaches in abnormality detection. Mostly used techniques focus uniquely on motion information, ignoring abnormality information due to variations of object pose. Optic flows may not even be powerful enough to characterize abnormality [9]. This makes them impervious to abnormalities that do not involve motion outliers. The descriptors, such as pixel change histograms or other traditional background subtraction operations, explicitly characterize abnormality. Therefore, a more complete representation, that uses both appearance and motion, seems viable. Boiman and Irani [10] use spatio-temporal patches and declare regions to detect abnormalities. Kratz and Nishino [11] propose to use spatio-temporal gradients. It is apparent that a combined approach that uses both motion and some form of appearance model results in a better abnormality detection.

In this paper, we use two different descriptors which comprise both motion and appearance features. Our selection is motivated by the excellent performance of such feature descriptors in human action recognition [12], [13]. These two approaches have not been employed previously for abnormality

978-1-4799-4598-6/14/$31.00 ©2014 IEEE

recognition. First: we use a frame sequence descriptor, which is a histogram of optic flow values and the silhouettes [12]. Combining these in a histograms gives a rich feature vector, that allow us to recognize abnormal activities. Second: we use dense trajectory based features aligned with histogram of gradients (HOG), histogram of optic flow (HOF) and motion boundary histogram (MBH) descriptors [13]. Given an annotated sequences of frames, normal and abnormal, we train a SVM classifier to classify a test frame as either normal or abnormal. Although we use SVM for the first method directly, we use bag of visual words model as an intermediate step for the second method before using SVM due to the huge number of features that cannot be handled directly.

This paper is organized as follows: section 2 gives an overview of the two methods, that we have used to create feature descriptors. Section 3 gives an brief introduction about the datasets and evaluation methodology. We present the experimental results in section 4 and we compare our results with other reported methods. Finally, we briefly discuss about our approaches and performance in section 5.

## II. Methodology

The problem that we address in this paper, in summary, is whether abnormal activities can be detected using spatio-temporal descriptors. We consider both pose and motion to detect abnormalities. We select two different spatio-temporal approaches for our experiments. First, We use a combination of optic flow value and the silhouettes-based features. Second, We use dense trajectory based features aligned with HOG, HOF and MBH descriptors which are more complex than the first approach.

### A. Silhouettes and Optic Flow based Features (SOF)

First step is extracting the image sequence from the $320 \times 240$ video. Then, we extract the silhouettes of characters (actors and objects) using background subtraction from each frame and generate the optic flow values. After that, we concatenate both optic flow values and histogrammed silhouette to produce the motion descriptor [12] as described in section 2.2. Finally, we classify the abnormal versus normal activities by using a SVM classifier.

Feature extraction process can be categorized into two stages: First, local features are extracted from each frame. Second, global features are found through activity sequence, comprising of several frames.

*1) Local Features:* In our system, a local feature is a histogram of the silhouette of the actors/objects and of the computed optic flow values. We use frames scaled to $320 \times 240$ for silhouette extraction and to resample the flow vectors. The optical flow vector field $F$ is first split into two scalar fields corresponding to the horizontal and vertical components of the flow, $Fx$ and $Fy$. To compute the optic flow values, we use Lucas-Kanade algorithm [14]. Each channel is smoothed by a median filter to reduce the effect of noise. Two real-valued channels $Fx$ and $Fy$ and the binary channel silhouette are the three channels which constitute the histogram. To get the silhouette, we use the background subtraction method used in Sheikh and Shah [15], due to the dynamic backgrounds in videos.

Each of these channels is histogrammed using the following technique: we divide the frame into $2 \times 2$ sub-windows and then each sub-window is divided into 18 pie slices covering 20 degrees each. These pie slices do not overlap and the center of the pie is in the center of the sub-window. The flow values of each channel are integrated over the domain of every slice. The result is a 72 $(2 \times 2 \times 18)$-dimensional histogram. By concatenating the histograms of all 3 channels we get a 216-dimensional frame descriptor [12].

*2) Motion Descriptor:* In creating the motion descriptor, the most important question is what the appropriate features are. Aforementioned feature extraction method can be adopted to capture very rich representations by incorporating static and dynamic features. That means, it can capture the local appearance and local motion of an object. It helps to detect abnormal and normal activities in the video frames.

Following Du Tran *et al.* [12], we too use 15 frames and split them into 3 blocks of 5 frames, each named as past, current and future. The frame descriptors of each block are stacked together into a 1080 $(72 \times 15)$ dimensional vector. The frame block descriptor is then projected onto the first $N$ principal components using PCA. Here the central idea of using PCA is as usual to reduce the dimensionality of the dataset. Then we keep first 50, 10, and 10 dimensions for the current, past, and future blocks respectively. Our idea is, give more emphasis to the nearby frames, as “local” motion gives better details than distant frames. Then the resulting 70-dimensional context descriptor is appended to the current frame descriptor to form the final 286-dimensional motion context descriptor.

### B. Dense Trajectory Based Features (DTF)

In this method, we first compute dense trajectories [13] and then create the trajectory aligned HOG, HOF and MBH descriptors. This is a complete representation of both motion and appearance features. This method is complex and time consuming compared to the SOF method. After creating these descriptors we use a standard bag of visual words approach followed by SVM to classify abnormal vs. normal activities in videos.

*1) Dense Trajectories:* First dense optical flow field is computed, and then points are tracked very densely for multiple spatial scales. To form a trajectory, these computed points of subsequents frames are concatenated. Following Wang *et al.* [13], we too limit the length of a trajectory to 15 frames to overcome the common problem called “drifitng” that arises in tracking. Fig. 1 shows the trajectories that we have calculated in our experiments.

*2) HOG, HOF and MBH Descriptors:* To get the motion information from the dense trajectories, within a space-time volume around the trajectory, descriptors can be computed. The size of the space-time volume is $N \times N$ pixels and 15 frames. Then, this volume is subdivided into a spatio-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$. Since motion and appearance features contribute to recognize abnormal and normal activities in video, it is important to get them from the computed trajectories. Among the existing such feature extraction methods, HOGHOF [16] and MBH [17] give excellent results in activity recognition in various datasets. HOG [18] mainly captures

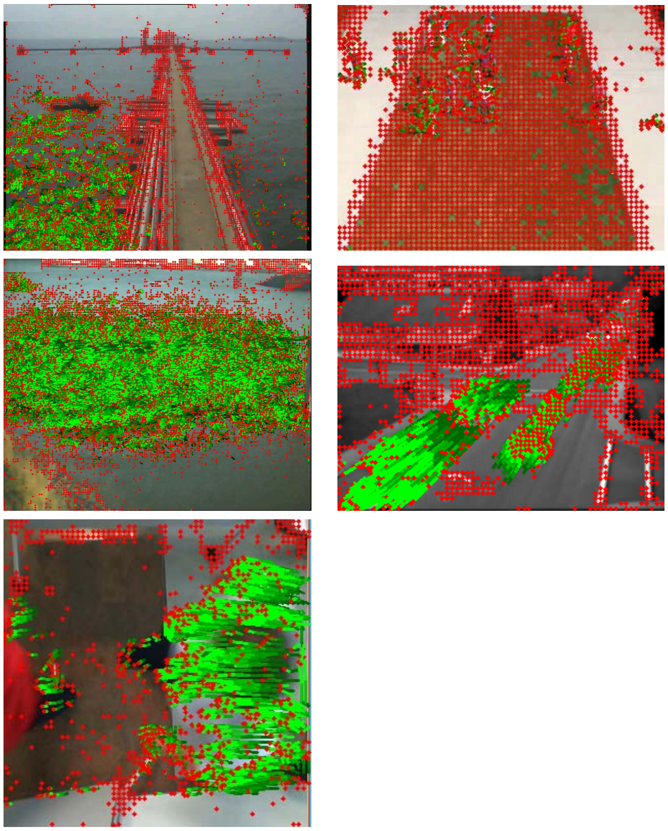

Fig. 1. Sample frames from dense trajectories of datasets that we have used. The datasets are Boat-Sea, Crowd Activity Scene 01, Boat-River, Traffic-Belleview and Airport-Wrong-Direction

static appearance features and HOF computes local motion information. HOF is mainly focus on absolute motion and it captures the motion information along with the camera motion. To overcome this problem Dalal *et al.* [17] proposed a method called MBH where derivatives are computed separately for the horizontal and vertical components of the optical flow. For HOG HOF and MBH, we create 8-bin histograms and normalize separately with $L_2$ norm. Finally, we have four separated descriptors, trajectory, HOG, HOF and MBH for each datasets.

## III. Datasets and Evaluation Methodology

### A. Description of Datasets

To test the effectiveness of the proposed algorithms, we systematically apply it to several datasets. The selected datsets are with both human and non-human abnormal activities. We select UMN dataset [19] for detecting crowd-scene abnormalities. Other datasets [20] have various types of abnormalities.

*UMN Dataset*: The UMN dataset has been collected by University of Minnesota [19] which consists of 3 different scenes of crowded escape events. The total frame number is 7740 (1450, 4415 and 2145 for scenes 1–3, respectively) at a $320 \times 240$ resolution. Fig. 2 shows some sample frames both abnormal and normal activities in the different scenes.

*Other Datasets*: We selected four more datasets with different abnormalities to evaluate our proposed methods. Fig. 3 shows the sample frames of all these four datasets both normal and abnormal activities.

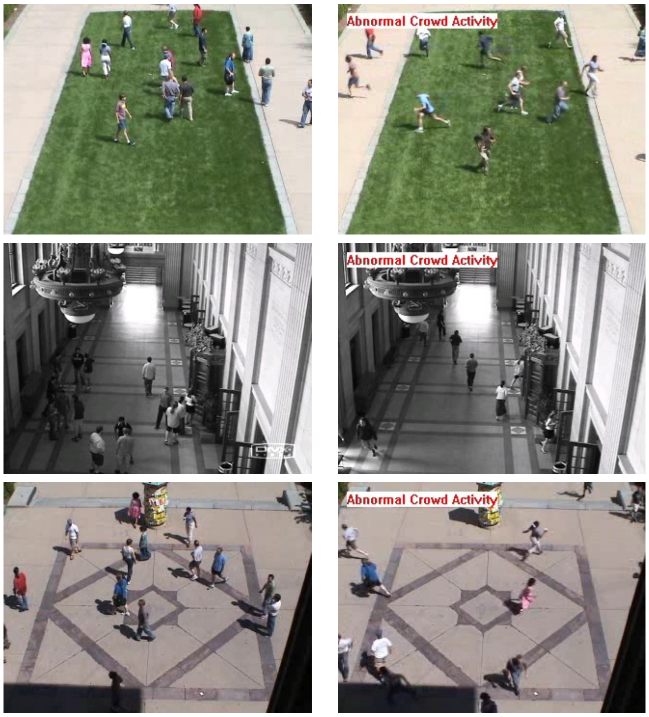

Fig. 2. Sample frames from some actions in UMN dataset. Left column: samples from the normal events; Right column: samples from the abnormal events. In the abnormal event people are suddenly deviating from the scene.

TABLE I. Description of the Boat-Sea, Boat-River, Traffic-Belleview and Airport-Wrong-Direction Datasets

| Dataset | Description | Abnormality |
|---|---|---|
| Boat-Sea | A sea-boat is passing by (motion on motion) Resolution: 720 × 576 | Boat movement |
| Boat-River | A boat is passing by on a river (motion on motion) Resolution: 720 × 576 | Boat movement |
| Traffic -Belleview | Cars are moving through and intersection. Resolution: 320 × 240 | Entering cars through from left to right |
| Airport -Wrong -Direction | Passengers are moving in a certain gate at the airport Resolution: 320 × 240 | A passenger moving in the wrong direction |

### B. Evaluation Methodology

To perform abnormal activity classification using the aforementioned SOF and DTF descriptors, a one-class SVM classifier was trained with labeled normal/abnormal descriptors. For DTF, before using a SVM classifier, we use standard bag of visual words method to create the codebook. We use half of the frames for training and rest for testing.

*1) Abnormal Activity Detection:* We used SVM classifier to solve this one-class problem with a low computational cost. To estimate the best classifier for our datasets, we carry out a grid search in the space of parameter $C$ and $\gamma$. Here, $C$ is the weight of error penalty and $\gamma$ determines the width of the RBF kernel. The appropriate SVM classifier is selected by the set of ($C$, $\gamma$) which maximizes the cross-validation rate in the space of search, which, in turn, increases the accuracy of the

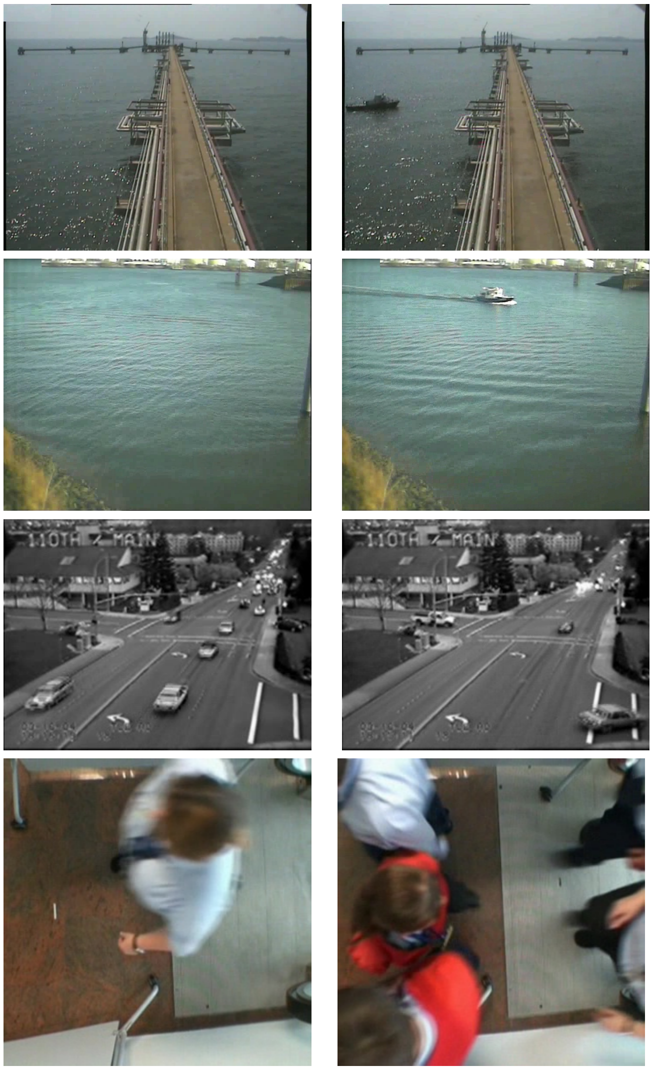

Fig. 3. Sample frames of titled Boat-Sea, Boat-River, Traffic-Belleview and Airport-Wrong-Direction datasets. Left column: samples from the normal events; Right column: samples from the abnormal events

results [21]. For the DTF method, we use a standard bag of visual words model before using SVM.

When using the standard bag of visual words model, we construct codebooks for each the descriptors (trajectory, HOG, HOF, MBH) as explained in the DTF approach, separately. We use a fixed number of visual word for our experiments, $4000$ which has given good results. We use $k$ means clustering with randomly selected $100,000$ features. It helps us to reduce the complexity of the feature descriptors. Descriptors are assigned to their closest vocabulary word using Euclidean distance. Then we use SVM, for each descriptor separately and get the average for each dataset.

## IV. Experimental Results and Discussion

### A. UMN Dataset

Table 2 shows that DTF method outperforms the state-of-the-art results and SOF method too performs comparably with other existing work. All the videos in this dataset exhibit behavior of a crowd in a panic situation at different places with different illumination levels. Both the methods successfully model the dynamics of the abnormal behavior, depending on the scene characteristics. Dense trajectories perform well with these three scenes with excellent results. In these videos, normal and abnormal frames differ significantly. When considering the SOF method, results for scene 2 are poor due to the illumination changes. Since, SOF considers both silhouette and optic flows, silhouette cannot perform well in such changes in the videos. It is difficult to compare the results with the existing methods due to the lack of information, such as, number of training examples, number of testing examples and computational time. However, dense trajectories perform excellently well in this dataset.

TABLE II. Recognition rates of abnormal activity using the UMN Dataset

| Method | Recognition rate |
|---|---|
| Ours (DTF), Scene 1 | **1.00** |
| Ours (DTF), Scene 2 | **1.00** |
| Ours (DTF), Scene 3 | **1.00** |
| Ours (SOF), Scene 1 | **0.96** |
| Ours (SOF), Scene 2 | **0.84** |
| Ours (SOF), Scene 3 | **0.95** |
| Cong *et al.* [8] | 0.97 |
| Cui *et al.* [22] | 0.95 |
| Mehran *et al.* [6], Social Force | 0.87 |
| Mehran *et al.* [6], Optic Flow | 0.81 |

### B. Other Datasets

We present the results of datasets, titled *Boat-Sea*, *Boat-River*, *Traffic-Belleview* and *Airport-Wrong-Direction* in Table 3.

TABLE III. Recognition rates (RR) of abnormal activity using the Boat-Sea, Boat-River, Traffic-Belleview and Airport-Wrong-Direction Datasets

| Dataset | DTF | SOF | Andrei *et al.* [23] |
|---|---|---|---|
| Boat-Sea | **1.00** | **0.97** | 1.00 |
| Boat-River | **1.00** | **1.00** | 1.00 |
| Traffic-Belleview | 0.6 | **1.00** | 0.9 |
| Airport-Wrong -Direction | 0.75 | **1.00** | |

According to the results of the Table 3, SOF method performs well in all four datasets and DTF method perform poorly in two datasets (Traffic-Belleview and Airport-Wrong-Direction). Frames in these two datasets may contain both abnormal and normal activities. DTF performs well in videos which contains normal and abnormal frames separately. In Traffic-Belleview dataset, abnormality is a part of its frame. For example, while cars are moving through an intersection (normal), cars traveling from left to right is the abnormality. Similarly, in Airport-Wrong-Direction dataset, abnormality is people moving in the wrong direction, while others are moving correctly. Dense trajectories generate a huge number of features and it is difficult to handle all. Random sub-sampling is one solution. This may cause bias towards normal classes associated with greater share of videos or, lengthy video sequences. These may have been the reasons for the poor performance with these two data sets. Except for those data sets, DTF outperforms the existing methods.

*Computational Time*

TABLE IV. COMPUTATIONAL TIME FOR THE TRAINING AND TESTING

| Method | Speed | Frame Size | Time |
|---|---|---|---|
| Ours | 2.0GHz | $320 \times 240$ | **15ms** |
| ANCI C [23] | 2.4GHz | $160 \times 120$ | 80ms |
| SSE2 [23] | 2.4GHz | $160 \times 120$ | 24ms |
| Mahadevan [9] | 3GHz | $160 \times 240$ | 2hrs |

According to the results in Table 4, SVM performs excellently well for both training and testing in terms of computational time.

## V. CONCLUSION

In this paper, we used two spatio-temporal descriptors, DTF and SOF, for recognizing abnormalities in videos. Spatio-temporal features, can recognize abnormal activities with high accuracy. DTF performs well in most of the videos, which have abnormal and normal frames separately. Due to limitation of the dense trajectories, it gives poor results when both normal and abnormal actions appear in the same frame. SOF, silhouette and optic flow based method performs well in all the datasets, but if illumination changes present, it tends to give poor results. To get the silhouette, we use a Bayesian modeling based approach that can handle dynamic backgrounds and it improves our results in the SOF method. In crowd videos, DTF outperforms the state-of-the-art methods. DTF can handle such abnormalities with near-perfect accuracy. Both motion and appearance features based descriptors perform well and additional trajectory shape allows DTF method to give better results in crowd dataset than SOF. SVM performs well in the both training and testing times when compared to other methods. Creating DTF descriptors is complex and it requires high performance hardware.

## REFERENCES

[1] Hua Zhong, Jianbo Shi, and Mirko Visontai, "Detecting unusual activity in video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, June–July 2004, vol. 2, pp. 819–826.

[2] Jaechul Kim and Kristen Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 2921–2928.

[3] Arslan Basharat, Alexei Gritai, and Mubarak Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008, pp. 1–8.

[4] Tianzhu Zhang, Hanqing Lu, and Stan Z. Li, "Learning semantic scene models by object classification and trajectory clustering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 1940–1947.

[5] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, March 2008.

[6] Ramin Mehran, Alexis Oyama, and Mubarak Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 935–942.

[7] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, and Iain McCowan, "Semi-supervised adapted hmms for unusual event detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, June 2005, vol. 1, pp. 611–618.

[8] Yang Cong, Junsong Yuan, and Ji Liu2, "Sparse reconstruction cost for abnormal event detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011, pp. 3449–3456.

[9] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010, pp. 1975 – 1981.

[10] Oren Boiman and Michal Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 1, no. 74, pp. 17–31, August 2007.

[11] Louis Kratz and Ko Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 1446 – 1453.

[12] Du Tran, Alexander Sorokin, and David Forsyth, "Human activity recognition with metric learning," in *Proceedings of the IEEE European Conference on Computer Vision*, Marseille, France, 2008, vol. Part I of *LNCS 5302*, pp. 549–562, Springer-Verlag Berlin Heidelberg.

[13] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011, pp. 3169–3176.

[14] Bruce D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stero vision," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, San Francisco, CA, 1981, vol. 2, pp. 674–679.

[15] Yaser Sheikh and Mubarak Shah, "Bayesian modelling of dynamic scenes for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1778 – 1792, November 2005.

[16] Ivan Laptev, Marcin Marszaek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008.

[17] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *Proceedings of the IEEE European Conference on Computer Vision*, Austria, May 2006, pp. 428–441.

[18] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, June 2005.

[19] "Abnormal crowd activity dataset of university of minnesota," available from http://mha.cs.umn.edu/movies/crowd-activity-all.avi.

[20] "An abnormal activity datasets," available from http://www.cse.yorku.ca/vision/research/anomalous-behaviour.

[21] Chia-Chih Chen and J. K. Aggarwal, "Recognizing human action from a far field of view," in *Proceedings of the IEEE Workshop on Motion and Video Computing*, Snowbird, Utah, December 2009, pp. 1–7.

[22] Xinyi Cui, Qingshan Liu, Mingchen Gao, and Dimitris N. Metaxas, "Abnormal detection using interaction energy potentials," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011, pp. 3161–3167.

[23] Andrei Zaharescu and Richard Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Proceedings of the IEEE European Conference on Computer Vision*, Crete, Greece, September 2010, vol. Part 1 of *LNCS 6311*, pp. 563–576.