DETECTING ABNORMAL BEHAVIORS IN SURVEILLANCE VIDEOS BASED ON FUZZY CLUSTERING AND MULTIPLE AUTO-ENCODERS

Zhengying Chen, Yonghong Tian, Wei Zeng, Tiejun Huang*

School of Electronics Engineering and Computer Science, Peking University, Beijing, P.R.China 100871
{cathychenchn, yhtian, weizeng, tjhuang}@pku.edu.cn

ABSTRACT

In this paper, we present a novel framework to detect abnormal behaviors in surveillance videos by using fuzzy clustering and multiple Auto-Encoders (FMAE). As detecting abnormal behaviors is often treated as an unsupervised task, how to describe normal patterns becomes the key point. Considering there are many types of normal behaviors in the daily life, we use the fuzzy clustering technique to roughly divide the training samples into several clusters so that each cluster stands for a normal pattern. Then we deploy multiple Auto-Encoders to estimate these different types of normal behaviors from weighted samples. When testing on an unknown video, our framework can predict whether it contains abnormal behaviors or not by summarizing the reconstruction cost through each Auto-Encoder. Since there are always lots of redundancies in the surveillance video, Auto-Encoder is a pretty good tool to capture common structures of normal video sequences automatically as well as estimate normal patterns. The experimental results show that our approach achieves good performance on three public video analysis datasets and statistically outperforms the state-of-the-art approaches under some scenes.

Index Terms— abnormal behaviors, video anomaly detection, fuzzy cluster, Auto-Encoder

1. INTRODUCTION

As security gains more and more importance in the modern daily life, surveillance cameras are broadly deployed. In this case, surveillance video analysis becomes a crucial task. Detecting abnormal actions or events in surveillance video is one of the most challenging analysis tasks, because the abnormal samples are usually quite rare and the concept of abnormalty is not well-defined on many occasions. Since it is difficult to list all the anomaly types as well as to obtain enough abnormal samples for building models, the anomaly detection task is usually regarded as an unsupervised problem. In this case, how to estimate normal patterns sufficiently becomes the key point.

*Corresponding to: Yonghong Tian, yhtian@pku.edu.cn

When facing daily life videos, it is inevitable that there are a large variety of normal behaviors, so using a single model is not enough to estimate the whole normal patterns. In this case, we propose a novel framework FMAE (Fuzzy clustering and Multiple Auto-Encoders) to estimate different types of normal patterns. Here, the Auto-Encoder is a special neural network whose input and output are the same, so that it can first encode the input through hidden layers and then decode it as the output. On one hand, common structures of the normal samples can be automatically learned by Auto-Encoders during the training phase. On the other hand, we can predict whether an unknown sample is abnormal or not according to its reconstruction cost through each Auto-Encoder and even indicate which normal type it may belong to.

Considering a sample may have features of different normal types, simply dividing the normal samples into several parts and building models for each part may not work well. To address this problem, our framework uses fuzzy clustering to measure the belonging degree of each sample towards each normal type. Thus, all the Auto-Encoders can be trained on the same set of samples with different sample weights. The sample's weight indicates the relevancy to a certain Auto-Encoder, and will affect the reconstruction cost as well. The larger the weight is, the more effect it will make to the reconstruction cost. In this way, all the Auto-Encoders share the whole normal samples, which takes full advantages of training information.

Our approach is inspired by the recent popular sparsity-based methods as well as the boosting algorithm. However, rather than building a codebook so as to describe all the normal patterns, we choose to use multiple weak learners to capture different normal structures so as to describe the normal behaviors more explicitly. We test the FMAE framework on the popular anomaly detection datasets, and the experimental results demonstrate that our approach yields good performances.

The rest of this paper is organized as follows. Section 2 reviews previous work on video anomaly detection. Section 3 describes the FMAE framework and the experimental results are shown in Section 4. Finally, Section 5 makes a brief conclusion.

2. RELATED WORK

Basically, there are two kinds of video anomaly detection tasks. One targets at pre-defined types of anomalies and can be regarded as a special action or event classification task. The other aims to detect un-defined anomalies and usually only has normal videos for training. Since we can neither list all types of anomalies nor get enough training samples for each type on most daily life occasions, we focus on the latter type of task in this paper.

Generally speaking, most of the existing unsupervised video anomaly detection approaches use the following pipeline: 1) Extract low-level visual features or get higher level descriptors with semantic meaning so as to represent the video content; 2) Estimate a normalcy model according to normal samples; 3) Apply anomaly detection by checking whether the testing data can fit the normalcy model or not.

While some approaches focus on combining spatial and temporal information into novel descriptors, others concentrate more on building different models to describe normal patterns.

2.1. Representations

In existing anomaly detection approaches, descriptors based on optical flow field seem to be the most popular features. Apart from some approaches using the optical flow directly or aggregating the statistical information like HOF [1, 2, 3, 4], there are also some approaches combining optical flow with other models. Shandong Wu et.al. [5] proposed the feature that uses chaotic invariants of Lagrangian particle trajectories, which can represent the trajectories in the crowded scenes well. Ramin Mehran et.al. [6] used the social force model to describe the motions, which takes the people's social interaction into consideration.

Features representing visual appearance information like corners, edges or gradients[7, 8] are also employed in some approaches. There are also some approaches tried to combine spatial and temporal features into a whole descriptor, so that abnormal motions as well as textures can both be detected. Vijay Mahadevan et.al. [9, 10] proposed the Mixture Dynamic Texture (MDT) model to represent the video content, which works well under crowded scenes.

However, all these descriptors need authors' strong background or priors in related fields, which limits their generalization in different anomaly detection tasks where background or priors are not always known. Instead, FMAE tries to reveal hidden patterns automatically by Auto-Encoders and yields good performance.

2.2. Models

In general, there are many models and their derivations which can be used to describe the normal behaviors, such as Gaussian mixture model (GMM)[5], Hidden Markov Model (HMM)[7], etc. Some approaches also apply the models in Natural Language Process (NLP) field to solve video anomaly detection tasks, such as topic models[11], Latent Dirichlet Allocation (LDA)[6], etc. Considering that videos contain both spatial and temporal information, there are also some approaches applying spatial-temporal models like Spatial-Temporal Markov Random Field (ST-MRF)[4, 12]. However, training these complicated models needs a large number of samples and is usually very time consuming.

Background subtraction algorithms and video parsing techniques can also be deployed to detect the anomalies[13, 14], for the abnormal behaviors always happen in the foreground field. Venkateshi Saligrama et.al. [3] proposed an approach based on local statistical aggregates, which relies on the assumption that the abnormal area is different from its neighboring fields.

In recent years, with the development of sparse coding theories, some approaches[1, 2, 8, 14] use a trained codebook to describe the normal behaviors and predict the anomaly according to the reconstruction cost. FMAE follows this idea, but relies on many simple Auto-Encoders rather than a huge codebook. This boosting-like approach is proven to make sense according to the good experimental results.

3. THE PROPOSED APPROACH

3.1. Framework

The framework of FMAE consists of a training phase and a testing phase.

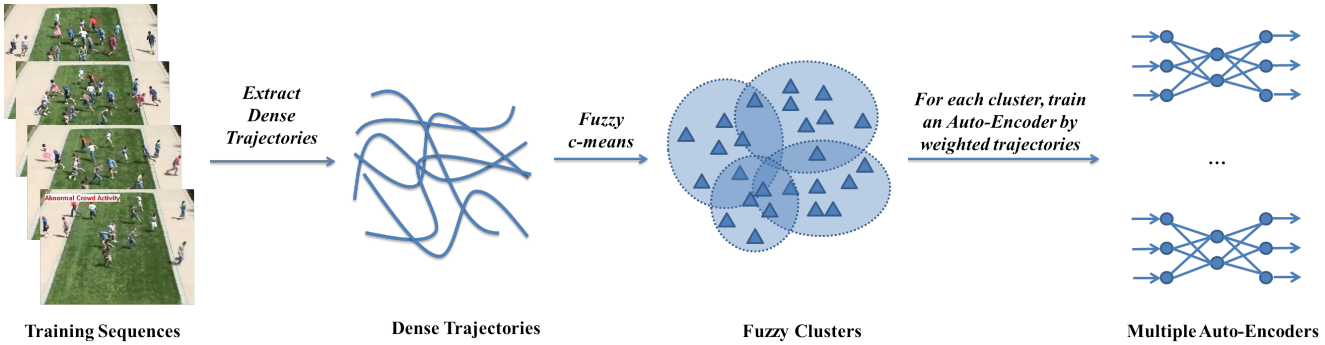
In the training phase, we use multiple Auto-Encoders to estimate normal patterns and the work flow is illustrated in Figure 1(a). Firstly, we extract the dense trajectories[15] of the training videos to obtain a trajectory pool. Then we distribute these trajectories into several clusters by fuzzy clustering techniques so that each cluster stands for a normal pattern. For each cluster, we use an Auto-Encoder to capture the common structure as well as to build a normalcy model.

During the testing phase, we predict an unknown patch's abnormal score by the trained Auto-Encoders and Figure 1(b) shows the work flow. When an unknown patch sequence comes, we extract its dense trajectories and calculate the reconstruction cost of each trajectory through each Auto-Encoder. The anomaly detection result can be inferred by summarizing these reconstruction costs.

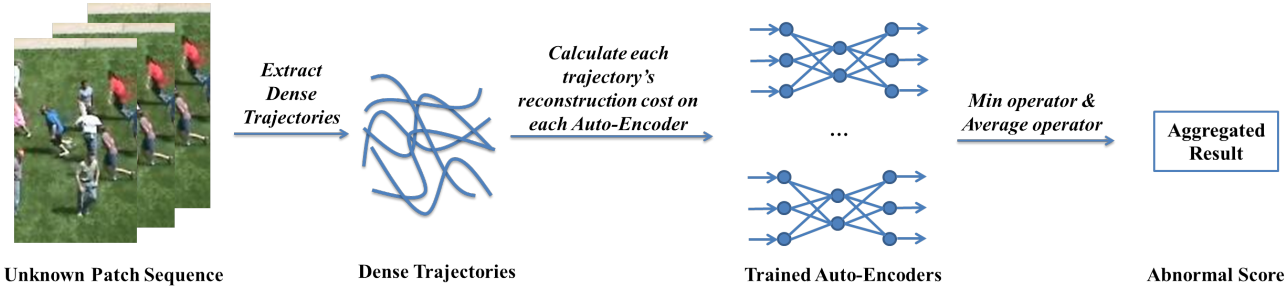
3.2. Estimate Normalcy Model

The first step of the training phase is to extract the dense trajectories[15] of training videos. Here we use a fixed length of trajectories and note it as L. The extracted trajectories are noted as $X \in R^{n \times p}$, where n is the number of trajectories and p is the dimension of the trajectory.

Then we do fuzzy clustering on X by Fuzzy c-means algorithm to get K clusters. The clustering procedure is



(a) The training phase of FMAE.



(b) The testing phase of FMAE.

Fig. 1. The framework of FMAE, including a training phase and a testing phase.

'fuzzy', indicating that each sample may belong to any cluster rather than a specific one. The probability is represented as a belongingness matrix $U \in R^{n \times K}$, where n is the number of trajectories, and K is the number of clusters. $U_{i,j}$ indicates the probability that the i^{th} trajectory belongs to the j^{th} cluster.

For each cluster, we train an Auto-Encoder to model the normal pattern it stands for. An Auto-Encoder is a special neural network whose input and output are the same[16]. Thus, the Auto-Encoder can be used as an unsupervised model, for it doesn't need any labels. The loss function of the j^{th} Auto-Encoder is defined as below:

$$L_j = \frac{1}{2} \sum_{i=1}^{n} c_{i,j} \cdot \|X^*_{i,j} - X_i\|_2^2 \tag{1}$$

where n is the number of trajectories, $c_{i,j}$ is the i^{th} trajectory's weight for the j^{th} cluster, X_i is the i^{th} trajectory, and $X^*_{i,j}$ is the reconstructed one according to the j^{th} Auto-Encoder.

The weight of trajectory $c_{i,j}$ can be obtained according to the belongingness matrix U as below, and satisfies $\sum_{k=1}^{n} c_{k,j} = n$.

$$c_{i,j} = U_{i,j} \cdot n / \sum_{k=1}^{n} U_{k,j} \tag{2}$$

After adding the regularization penalty to avoid over-fitting, we can get the cost function as follows.

$$J_{j,k} = L_j + \frac{\lambda}{2} \sum_{s=1}^{p_k} \|w_{j,k}\|_2^2 \tag{3}$$

where $J_{j,k}$ is the cost function of the k^{th} layer in the j^{th} Auto-Encoder, $w_{j,k}$ is the model parameter, and p_k is the number of parameters. The model parameter $w_{j,k}$ can be simply trained through the Back Propagation algorithm.

3.3. Anomaly Detection

When an unknown patch sequence comes, we extract its dense trajectories and calculate the reconstruction cost of each trajectory through each Auto-Encoder.

To evaluate a trajectory's abnormal score, we use the following criterion:

$$RC_i = min\{\|Y^*_{i,j} - Y_i\|_2^2\} \tag{4}$$

where RC_i is the abnormal score of the i^{th} trajectory, $Y^*_{i,j}$ is the reconstructed result of this trajectory according to the j^{th} Auto-Encoder, and Y_i is the i^{th} original trajectory.

As for a patch, its abnormal score is the average reconstruction cost of all the trajectories in it:

$$score_P = \frac{1}{N_k} \sum_{k} RC_k, \forall k \in P \tag{5}$$

where $score_P$ is the abnormal score of a patch P, and N_k is the number of trajectories in this patch. When a trajectory goes through this patch, we regard it as in this patch.

If the abnormal score is higher than a threshold th, then we consider this patch contains abnormal behaviors. The parameter th can be trained through a validation set or be chosen manually. The patch size indicates the scale of detected behaviors, i.e. small patches reflect local behaviors while large patches even the whole frame indicate global behaviors.

4. EXPERIMENTS

4.1. Protocal

We did experiments on three popular video analysis datasets, including the UMN dataset[1], the subway dataset[2] and the QMUL junction dataset[3]. The training sequences only contain normal events while the testing ones contain both normal and abnormal behaviors.

In the experiments, the length of trajectories L was fixed at 15, and the cluster number K during the fuzzy clustering phase was fixed at 100. Using multi-scale length of trajectories or more clusters may help to improve the performance, but our experimental results demonstrate that the above setting is just good enough considering the time cost.

We deployed three-layer neural networks as the Auto-Encoder, i.e. the Auto-Encoder only has one hidden layer. The number of hidden layers in the Auto-Encoder can be extended so as to construct a deeper structure and capture the patterns better, but here we choose one-hidden-layer Auto-Encoders for simplicity.

4.2. Experimental Results

4.2.1. UMN Dataset

The UMN dataset consists of two outdoor scenes and one indoor scene, and contains several crowd rapid escape behaviors under each scene. The total number of frames is 7,739 and the resolution is 320 by 240. For each scene, we use 500 to 1,600 normal frames for training, and the rest frames are used for testing.

We evaluate the performance of our approach by calculating the area under the ROC (Reveiver Operating Characteristic) curve. The ROC curve shows the result of true positive rate(TPR) vs. false positive rate(FTR). TPR is the ratio of true positive frames to positive frames, and FPR is the ratio of false positive frames to negative frames. The area under the ROC curve illustrates the accuracy as well as robustness of an approach, i.e. the larger the area is, the better the performance is. The experimental result on the UMN dataset is shown in Table 1. Figure 2 gives a snapshot of this dataset, including both normal and abnormal behaviors under each scene.

Approach	area under ROC
SRC[1]	0.978
local statistical aggregates[3]	0.985
MDT[9]	0.995
Social Force[6]	0.96
Chaotic Invariants[5]	0.99
FMAE	0.97

Table 1. Experimental results on the UMN dataset.

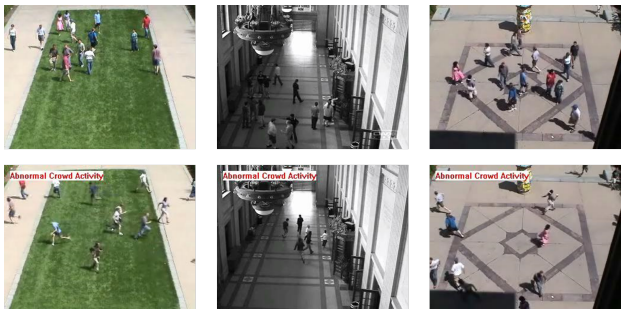


Fig. 2. UMN dataset: three scenes with crowd rapid escape behaviors. The frames in the top row show normal crowd behaviors, and the ones in the bottom row show abnormal behaviors.

Relatively speaking, the UMN dataset is quite simple compared with other datasets, so most of current anomaly detection approaches perform very well on this dataset. FMAE reaches a comparable performance with other state-of-the-art approaches[1, 3, 9, 6, 5].

4.2.2. Subway Dataset

The subway dataset consists of two scenes: entrance and exit. The entrance video is about 1.5 hours long with 144,249 frames in total, while the exit video is about half an hour long with 64,901 frames. The resolution of these videos is 512 by 384, and the videos are all in gray scale. We use 3,800 normal frames for training under the entrance scene and 2,100 frames under the exit scene, which uses much smaller training set than that in other approaches.

Subway Entrance

The abnormal behaviors under the subway entrance scene can be approximately divided into four types, including walking in a wrong direction, entering without payment, loitering around the gate and irregular interactions like calling subway customer service.

The statistical results on the subway entrance dataset are shown in Table 2. Figure 3 shows several detected abnormal behaviors for a few frames.

[1] http://mha.cs.umn.edu/proj_events.shtml
[2] From the author of [17]
[3] http://www.eecs.qmul.ac.uk/ ccloy/downloads_qmul_junction.html

Approach	WD	NP	LT	II	False Alarm
Ground Truth	26	13	14	4	/
Fast SRC[8]	25	9	14	4	5
Dynamic SC[2]	26	7	13	4	4
MPPCA[4]	24	7	14	3	3
FMAE	26	10	13	4	4

Table 2. Experimental results on the subway entrance dataset. WD: Wrong Direction; NP: No Payment; LT: Loitering; II: Irregular Interactions.

Fig. 3. Subway entrance dataset: top-left frame shows a man loitering at the gate; top-right frame shows a man trying to enter the gate without payment; bottom-left frame shows a man trying to exit from the entrance gate; bottom-right frame shows a woman and a service man talking at the gate.

Approach	WD	LT	MISC	False Alarm
Ground Truth	9	3	7	/
Fast SRC[8]	9	3	7	2
Dynamic SC[2]	9	3	7	2
MPPCA[4]	9	3	6	0
FMAE	9	3	6	3

Table 3. Experimental results on the subway exit dataset. WD: Wrong Direction; LT: Loitering; MISC: misc.

Fig. 4. Subway exit dataset: the left frame shows a man walking from the exit to the platform which is in a wrong direction; the right frame shows two men loitering on the platform while a service man standing still and watching them.

FMAE yields a high detection rate as well as a low false alarm rate on this dataset, and outperforms the state-of-the-art approaches on detecting wrong direction and no payment behaviors.

Subway Exit

The subway exit scene is relatively simpler which contains fewer types of abnormal or interesting behaviors, including walking in a wrong direction, loitering around the exit gate, etc.

The statistical results on the subway exit dataset are shown in Table 3. Figure 4 shows several detected abnormal behavior results for a few frames.

4.2.3. QMUL junction dataset

We also tested FMAE by detecting non-human behaviors like traffic events on the QMUL junction dataset. The QMUL junction dataset features a public road junction controlled by traffic lights and dominated by four traffic flows[18]. The video is about 1 hour long with 90,000 frames, and the resolution is 360 by 288. There are several types of abnormal behaviors, including illegal U-turn, improper lane of traffic, jaywalking, etc.

We choose 1,500 frames to train the model, and Figure 5 gives a snapshot including both normal and abnormal behaviors. The detecting results show that our approach can be applied to detect not only human abnormal behaviors but also non-human events like traffics.

5. CONCLUSION

In this paper, we present a novel framework to detect abnormal behaviors in surveillance videos by using fuzzy clustering techniques and multiple Auto-Encoders (FMAE). The Auto-Encoder is able to capture common structures of normal video sequences automatically, and multiple Auto-Encoders can efficiently estimate different types of normal patterns in the daily life. Fuzzy clustering assures that the Auto-Encoders take full advantages of training information, so we don't need too many samples. The experimental results show that our approach yields good performance on public video analysis datasets and outperforms the state-of-the-art approaches under some scenes.

Acknowledgement

This work is partially supported by grants from the National Basic Research Program of China under grant 2015CB351806, and the National Natural Science Foundation of China under contract No.61390515 and No.61471042.

Fig. 5. QMUL junction dataset: first two rows show normal traffic flows; bottom-left frame shows a detected illegal U-turn car; bottom-right frame shows a jaywalking passer-by.

6. REFERENCES

[1] Yang Cong, Junsong Yuan, and Ji Liu, "Sparse reconstruction cost for abnormal event detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3449–3456.

[2] Bin Zhao, Li Fei-Fei, and Eric P Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3313–3320.

[3] Venkatesh Saligrama and Zhu Chen, "Video anomaly detection based on local statistical aggregates," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2112–2119.

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

[5] Shandong Wu, Brian E Moore, and Mubarak Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2054–2060.

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

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

[8] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of International Conference on Computer Vision*. IEEE, 2013, pp. 2720–2727.

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

[10] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[11] Tom SF Haines and Tao Xiang, "Delta-dual hierarchical dirichlet processes: A pragmatic abnormal behaviour detector," in *Proceedings of International Conference on Computer Vision*. IEEE, 2011, pp. 2198–2205.

[12] Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2458–2465.

[13] Borislav Antic and Björn Ommer, "Video parsing for abnormality detection," in *Proceedings of International Conference on Computer Vision*. IEEE, 2011, pp. 2415–2422.

[14] Mehrsan Javan Roshtkhari and Martin D Levine, "Online dominant and anomalous behavior detection in videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 2611–2618.

[15] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *Proceedings of International Conference on Computer Vision*. IEEE, 2013, pp. 3551–3558.

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[17] 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, 2008.

[18] Chen Change Loy, Tao Xiang, and Shaogang Gong, "Modelling multi-object activity by gaussian processes.," in *Proceedings of British Machine Vision Conference*, 2009, pp. 1–11.