Anomaly Detection in Crowd Scene via Online Learning

Dandan Ma[1,3], Qi Wang[2,*], Yuan Yuan[1]

[1]Center for OPTical IMagery Analysis and Learning (OPTIMAL),
State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China
[2]Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University,
Xi'an 710072, Shaanxi, P. R. China
[3]University of the Chinese Academy of Sciences, 19A Yuquanlu, Beijing, 100049, P. R. China
madandan@opt.cn; crabwq@nwpu.edu.cn; yuany@opt.ac.cn

ABSTRACT

Anomaly detection in crowd scene has attracted an increasing attention in video surveillance, but a precise detection still remains a challenge. This paper presents a novel online learning method to automatically detect abnormal behaviors in crowd scene. Our focus is mainly on the deviation between the real motion and the predicted one. Through online defining experts, analyzing their motions, and dynamically updating the learned model, anomaly can be identified by the final expert joint decision. The outputs are represented as the anomaly probability of an examined frame. Compared with most of existing methods, the proposed one needs neither tracking each individual straight to the end nor requires any complex training procedure. We test the proposed method on public datasets, and the results show its effectiveness.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, Experimentation, Performance

Keywords

Anomaly detection, crowd scene, object tracking, motion estimation, online learning.

1. INTRODUCTION

Anomaly detection on public places has become a hot topic in the computer vision community, due to the increasing security awareness. When some vicious incidents occur, abnormal behavior detection can be of great help to event investigation and forensics. However, in particular for the crowd scene, the detection of abnormal behaviors is a herculean task because of the severe clutter and high object density [13]. Therefore, developing an effective method to automatically detect anomalies in crowd scene still needs further research.

In recent years, a wealth of work have been committed to the crowd abnormal detection. Throughout the literatures, existing methods can be roughly divided into two categories: 1) Tracking based, which relies on the analysis of the individual trajectories, and 2) Model based, which directly models the activity patterns.

Tracking based methods [13, 5, 3] are the conventional approaches. In these methods, anomaly will be identified when trajectories significantly deviate from the learned normal ones in terms of certain measures [3]. Though tracking has been researched a lot in the field of visual surveillance, to track every individual in the crowd scene is still a challenging task [10, 14]. This leads to the difficulty of analyzing trajectories in anomaly detection.

Due to the deficiency of tracking individuals in a crowd scene, the model based methods become more popular recently. Most of these methods refer to the trained normal events and classify those do not conform to them as abnormal ones. For example, Mehran *et al.* [9] model the normal crowd behavior by using the social force model, which investigates the particle motion dynamics. Zhang *et al.* [15] further consider the disorder and congestion attributes on the basis of social force model [9]. Mahadevan *et al.* [8] utilize the mixtures of dynamic textures to model the normal crowd behavior. In spite of the popularity of this type of approaches, there is still a major bottleneck that cannot be ignored. As most of these methods need a rigorously training stage to learn the model parameters, they rely heavily on the availability of a large number of labeled training data. However, it is difficult to strictly satisfy this requirement in practice.

Taking fully into account the obstacles faced by the two kinds of anomaly detection approaches, a novel online learning method is proposed in this work. It's worth noting that we define the sudden motion change as the abnormal behavior with respect to human perception.

Our major *contributions* are threefold. 1) An online learning framework is presented to dynamically update the selected objects without a long-term training process. 2) An adaptive selection of objects is formulated to avoid the need of tracking a fixed target from the beginning to the end. 3) A joint decision strategy is adopted to tactfully combine the

ICIMCS'14, July 10–12, 2014, Xiamen, Fujian, China.

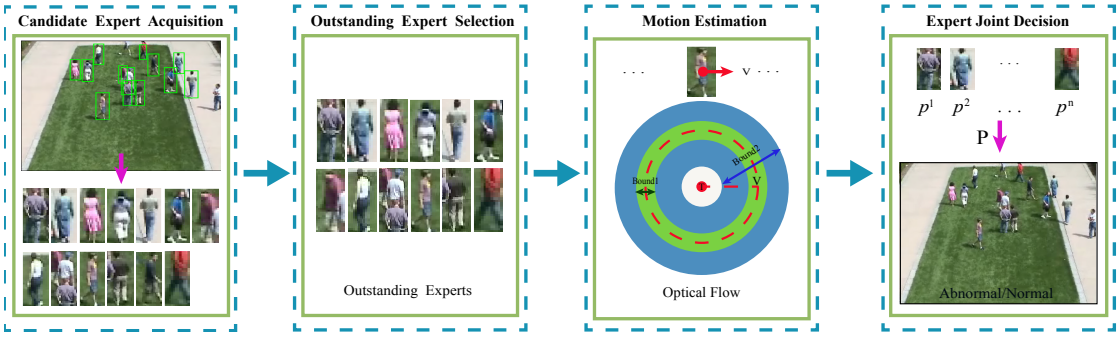

Figure 1: The flow chart of the proposed method for abnormal event detection in crowd scene.

advantages of the selected objects in together to ensure the accuracy of detection.

We continue this paper as follows. In Section 2, we introduce the proposed method in details. In Section 3, experimental results are reported and discussed. Finally, conclusions are made in Section 4.

2. METHOD

In this paper, we first define each individual obtained from the pedestrian detector as a candidate expert. Then the candidates well tracked between two adjacent frames are selected as the outstanding experts, which contribute to the following abnormal decision. After that, the motions of the outstanding experts can be calculated and compared with the predicted ones by analyzing their historical records. At last, the abnormal probability of the examined frame is jointly made by estimating all the motion deviations of the outstanding experts. The flowchart of the proposed method is illustrated in Fig. 1.

2.1 Candidate Expert Acquisition

Candidate expert acquisition aims at making preparations for the further selection of the well tracked ones, which will participate in analyzing the existence of abnormal events. In this work, for one examined frame of a given video sequence, the individuals detected by a pedestrian detector [6] are defined as the candidate experts. Then the position information of the bounding boxes corresponding to the detected individuals can be recorded. To be formal, the i^{th} frame of the input video is denoted as f_t and the position information of the i^{th} candidate expert in f_t is P_t^i $(i = 1, ..., N_{ce}^t$, where N_{ce}^t is the total number of the candidates in $f_t)$.

2.2 Outstanding Expert Selection

In order to obtain sufficient information of the entire crowd, the ideal solution is to synthetically analyze every individual's movement. However, as mentioned before, for a crowd, especially a highly dense one, it's still full of challenges to track each individual reliably [9]. What is worse, low resolution video frame makes tracking more like a tough nut to crack. Therefore, avoiding exhaustively tracking is an informed choice, which can make anomaly detection in the crowd scene more efficient.

As is well known that individuals compose the crowd, each of which is considered as an element of the crowd collection. For one individual in the crowd, its behavior is affected by others and similarly, the others are influenced by the individual as well. In other words, the individuals's behaviors to a certain extent can reflect the group movement trend. Therefore, we can select some representative ones to infer the motion state of the crowd by synthetical analysis. For this purpose, the candidates in current frame that can be well associated with the ones in the previous frame are defined as the outstanding experts. A diagram of the selection procedure is shown in Fig. 2. Assume the individuals with the bounding boxes are the experts obtained in f_{t-1}, and we aim to find the outstanding experts in f_t. Applying the KNN search, the five nearest individuals (red rectangles) of the examined candidate expert in f_t are obtained. Then using distance information, color histograms and HOG histograms, three similarity scores can be computed respectively. The individual with the maximum joint score is finally associated with the examined expert. The associated expert in f_t is the desired outstanding expert.

2.2.1 *Feature Description*

In the following, we describe the strategy to characterize each candidate expert, which simultaneously adopts object position, color histogram and HOG feature [7, 11]. These features are simple but effective to associate the detected individuals between two adjacent frames.

The first step is to extract the position of each candidate expert. After acquiring the candidate experts by using the pedestrian detection algorithm, the position information $P_t^i(x, w, y, h)$ of a detected target is obtained, where x and y denote the coordinates of the upper left corner of the bounding box, and w and h denote its width and height. In order to make it easier to understand, this work employs the center position $P_t^i(x_c, y_c)$ where $x_c = x + w/2$ and $y_c = y + h/2$. This central information is then used to characterize the distance of individuals between f_t and f_{t-1}.

The second step is to calculate the color histogram, which has been widely used to characterize images. Color histogram has the property of low complexity for feature description. Besides, this feature also has good robustness to noise as well as local image transformations. In this work, RGB space is chosen because it has been proved that RGB exhibits eminent performance with best average score. For RGB images, we allocate 8 bins for each color channel, yielding thus a color histogram vector with length 512.

The third step is to calculate the HOG feature, which is based on the consideration of the following problem. When the candidate experts from the same frame are similar to each other, color features may have lost their identification. Therefore, to remedy that, HOG feature is added to describe image patches. HOG operates on relatively small image regions, which makes itself keep good invariance to geometric

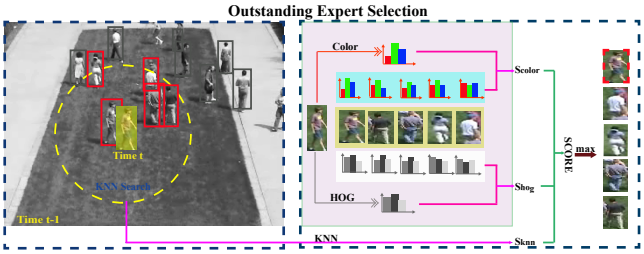

Figure 2: Illustration of outstanding expert selection.

and illuminative transformation. Besides, this feature can represent not just horizontal and vertical orientations but also edge information. All of these advantages make HOG suitable to characterize pedestrians.

2.2.2 *Similarity Measurement*

The outstanding experts are defined as the well associated candidates from two adjacent frames. The criteria includes one spatial distance and two feature distances.

For a candidate in f_t, we search for its possible associated correspondence from 5 adjacent neighbors in f_{t-1} based on the spatial distance. Then the search returns the adjacent distant degree $D_{knn}(c_1, c_2)$ with respect to the Euclidean distance, where c_1 and c_2 denote the comparing candidate experts from two adjacent frames.

According to the returned 5 results, the distances in feature space concerning color histogram and HOG feature are measured as well. A number of methods are optional for similarity analysis between two histograms. Among these methods the employed Bhattacharyya coefficient is one of the most common ones, which defines the distance between two histograms as

$$d(h_1, h_2) = \sqrt{1 - \sum_{b=1}^{n} \sqrt{h_1(b) \cdot h_2(b)}}, \tag{1}$$

where h_1 and h_2 are two normalized histograms and n denotes the total number of histogram bins. Using Eq. 1, two distances D_{color} and D_{hog} are computed respectively, where D_{color} is the distance between two color histograms, and D_{hog} is the distance between two HOG vectors.

With the obtained three distances, a final similarity score is calculated. To be specific, distance value is converted into similarity score with a Gaussian function

$$S(x, y) = e^{-\frac{D(x,y)^2}{2\sigma^2}}, \tag{2}$$

where $D(x, y)$ is the distance between two features x and y, and σ denotes the bandwidth of the Gaussian function which takes 0.3 in this work. Therefore, when putting the three distances indicated above into the distance-score converting formula, we get the similarity scores S_{knn}, S_{color} and S_{hog} respectively. Then, the comprehensive evaluation score $S(c_1, c_2)$ is generated by combining the three ones as follows:

$$S(c_1, c_2) = S_{knn}(c_1, c_2) \cdot S_{color}(c_1, c_2) \cdot S_{hog}(c_1, c_2). \tag{3}$$

Using this strategy, for each candidate expert, the most relevant target with the highest comprehensive evaluation score is picked out. Besides, a score threshold λ is also set in order to ensure the accuracy of tracking. This is because if the candidate has been occluded seriously or left the scene, there might be no truly matched candidate expert in the KNN search range. Therefore, the search range is further corrected by the constraint condition $S(c_1, c_2) > \lambda$. In this work, λ is intuitively set as 0.95.

After the selection process, N_{oe}^t individuals are picked out as the outstanding experts from a total of N_{ce}^t candidates.

2.3 Motion Estimation

As this work concentrates on the sudden motion change, each outstanding expert needs to estimate its motion pattern. The aforementioned association process cannot ensure a precise motion estimation because if the individual does not move but the limbs have large movement, the target association will indicate no motion change. Actually, this might be a sign of abnormality. In order to get a more precise motion estimation for further abnormal examination, we employ the optical flow technique to calculate each pixel's motion direction and magnitude. For each pixel, we use V_x and V_y to denote its horizontal and vertical velocity and the l_2 norm of (V_x, V_y) denote the motion magnitude. For an outstanding expert, its motion covers an rectangular patch containing lots of pixels. Therefore, the average magnitude of the whole patch is regarded as the representative motion.

2.4 Expert Joint Decision

After getting the motion information, we need to judge whether an abnormal behavior occurs for an examined frame. The result is expressed as an abnormal probability in view of each expert's decision. When finishing the final joint judgment, all the model parameters are online updated.

2.4.1 *Criterions*

The abnormal probability is estimated based on the deviation between the real motion and the predicted one. Therefore, these two motions should be determined at first. When an examined frame f_t needs to be judged, its representative motion $M_r^{i,t}$ of the i^{th} outstanding expert can be calculated first. For the determination of the predicted one, we need to make use of the historical information. In this work we don't manage to track each target from the beginning to the end in order to obtain its motions. Instead, only the outstanding experts, which are also named as *deciders*, are considered. We calculate the mean of the historical motions $M_h^{i,t}$ to represent the predicted one $M_p^{i,t}$. At the same time, the standard deviation σ of $M_h^{i,t}$ is computed as well.

In order to analyze the deviation degree, we need to calculate two kinds of motion ranges. According to the probability and statistics theory, the three-sigma principle is applied in this work to estimate them. Therefore, they are set by the following formula respectively:

$$\mathbf{Bound_1} = [M_p^{i,t} - \sigma, M_p^{i,t} + \sigma], \tag{4}$$

$$\mathbf{Bound_2} = [M_p^{i,t} - 3\sigma, M_p^{i,t} + 3\sigma], \tag{5}$$

where $\mathbf{Bound_1}$ is the predicted normal moving range , which means that the real motion $M_r^{i,t}$ should be within it, and $\mathbf{Bound_2}$ is the observable moving range, which means that the anomaly absolutely occur if $M_r^{i,t}$ goes beyond it. When $M_r^{i,t}$ is within $\mathbf{Bound_2}$ but out of $\mathbf{Bound_1}$, the object is probably abnormal in different degree.

By the above definition, the i^{th} decider can make its de-

cision about the abnormal probability as follows:

$$p_t^i = \begin{cases} 0.5 - 0.5\frac{\left\|M_r^{i,t} - M_p^{i,t}\right\|_2}{\sigma+\varepsilon}, & M_r^{i,t} \in \mathbf{Bound_1}; \\ 0.5 + 0.5\frac{\left\|M_r^{i,t} - M_p^{i,t}\right\|_2 - \sigma}{2\sigma+\varepsilon}, & M_r^{i,t} \in \mathbf{Bound_2} - \mathbf{Bound_1}; \\ 1, & M_r^{i,t} \cap \mathbf{Bound_2} = \emptyset, \end{cases} \tag{6}$$

where p_t^i denotes the decision of the i^{th} decider in f_t, and it reflects the abnormal probability from the i^{th} expert's point of view. The parameter ε is a small constant to make sure the value of the denominator is not zero. The larger p_t^i is, the more abnormal degree the examined frame has.

Based on the above equation, each decider has a judgement. Considering their different abilities to make a reliable decision, the deciders are weighted in accordance with their performances in the past. The final joint decision is expressed as

$$P_t = \sum_{i=1}^{N_{oe}^t} w_t^i p_t^i, \tag{7}$$

where w_t^i is the weight of the i^{th} decider, and P_t is the final abnormal probability of the examined frame f_t.

The difference between each individual's decision and joint decision is defined as a loss parameter which measures the correctness of each expert's decision. It is specified as

$$l_t^i = \left\|P_t - p_t^i\right\|_2. \tag{8}$$

In order to evaluate the ability of each decider to make decisions fairly, the cumulative loss L_t^i of the i^{th} decider in f_t is computed as follows

$$L_t^i = L_{t-1}^i + l_t^i. \tag{9}$$

At the same time, the weight of each expert is calculated based on it

$$w_t^i = 1/(N_{oe}^{t-1})e^{-\eta L_{t-1}^i}, \tag{10}$$

where η is a learning rate parameter, and is set to $\eta = 0.4$ empirically in this work.

2.4.2 *Parameter Update*

This work updates parameters at each frame in order to make a dynamic analysis and realize online learning. As mentioned above, only the outstanding experts can take part in the decision process. For the candidates that have not been selected as the outstanding ones, we define them as the new experts. They will be initialized to participate in the next frame. According to the joint decision of f_t, the motion information will be updated as follows:

$$\begin{cases} M_h^{i,t} = M_h^{i,t-1} \cup M_r^{i,t}, & i = 1, ..., N_{oe}^t, P_t \le \beta; \\ M_h^{j,t} = \frac{1}{N_{oe}^t}\sum_{i=1}^{N_{oe}^t} M_r^{i,t}, & j = N_{oe}^t + 1, ...N_{ce}^t, P_t \le \beta; \\ M_h^{i,t} = M_h^{i,t-1}, & i = 1, ..., N_{oe}^t, P_t > \beta; \\ M_h^{j,t} = \frac{1}{N_{oe}^{t-1}}\sum_{i=1}^{N_{oe}^t} M_r^{i,t-1}, & j = N_{oe}^t + 1, ...N_{ce}^t, P_t > \beta, \end{cases} \tag{11}$$

where j is the index of the new experts and β is a threshold. When the examined frame is identified as normality, the motion information is updated timely, but not vice versa. In our case, β is 0.5.

The weight and loss parameter are updated as follows:

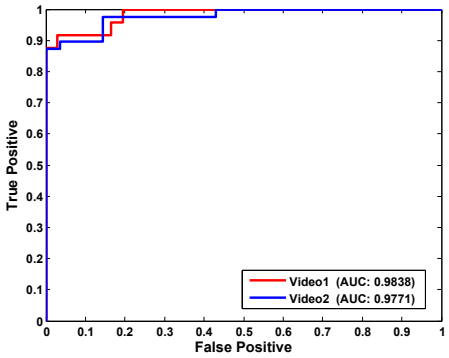

Figure 3: ROC of the proposed method on the PETS 2009 videos.

$$w_t^j = \frac{1}{N_{ce}^t}, \ L_t^j = 0, \qquad j = N_{oe}^t + 1, ..., N_{ce}^t, \tag{12}$$

$$w_t^i = \frac{N_{oe}^t}{N_{ce}^t}\frac{w_{t-1}^i}{\sum_{i=1}^{N_{oe}^t} w_{t-1}^i}, \qquad i = 1, ..., N_{oe}^t. \tag{13}$$

Using Eq. 12, the new experts are initialized. After that, all of the outstanding experts are reweighed using Eq. 13.

3. EXPERIMENTS

To validate the effectiveness of the proposed method, we test it on two popular datasets, the PETS 2009 dataset [1] and the UMN dataset [2]. Details are introduced in the following.

3.1 PETS 2009 Dataset

Two short videos of the PETS 2009, respectively depicting people from walking to running and people from assembling to escaping, are used in this work. To the best of our knowledge, only Wang *et al.* [12] conducted experiments on the first video of this dataset. Since their method needs a training process, the video is too short to be tested directly. So they employed another video to fulfill the training stage. Nevertheless, the proposed one can directly test on such short videos without any assistances. This is because the online learning can capture the objects' motions after a brief time, which makes our method more adaptive.

The ROC curves of the proposed method are illustrated in Fig. 3, and the AUCs of them are 0.9838 and 0.9771, respectively. This is highly effective because this dataset is very difficult. In order to fairly compare with [12] which computes the accuracy as its evaluation metric, we also compute the same kind of value. The result is 82.18% for [12] and 87.50% for ours, which further demonstrates superior performance of the proposed method in this work.

3.2 The UMN Dataset

The UMN dataset is provided by University of Minnesota, and consists of videos of 11 different scenarios for an escaping event. The videos comprise 3 different scenes, including lawn, indoor and square. All these videos start with an initial normal behavior and end with abnormal ones of people suddenly evacuating.

The proposed method is compared with four state-of-the-art competitors including the Social force model [9], the pure Optical flow model [9], the Streakline Potentials model [15]

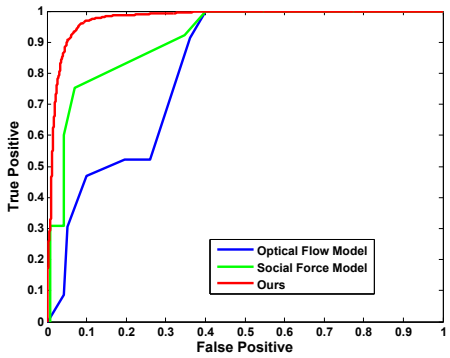

Figure 4: ROCs for abnormal detection on the UMN dataset.

and the Sparse reconstruction model [4]. Their results are directly cited from their original works. In Fig. 4, the ROC curves manifest that the proposed method outperforms the Social force model and Optical flow model greatly. Table 1 provides the results of quantitative comparison of AUC. It can be found that our method is also superior to the Streakline Potentials method. When it comes to the comparison with the Sparse method, the results of the three different scenes are estimated respectively. For the lawn and the indoor scenes, the proposed one is slightly inferior, but for the square scene, it's significantly better. As a whole, our method is also competitive with the Sparse method.

To sum up, the proposed method outperforms the state-of-the-art competitors for three reasons. First, the online learning characteristic ensures that the tiny change with previous observations can be observed immediately. This strategy avoids tedious training and is very adaptive. Second, only the limited number of experts should be tracked in the whole procedure, whose information is stable and reliable. Third, the joint decision combines different decisive abilities and is very robust.

4. CONCLUSION

In this paper, we propose a novel online learning framework for anomaly detection in the crowd scene. In our work, the adaptive selection of experts and the dynamical update of the model are adopted to avoid an exhausting training stage and a long-term tracking as well. At the same time, it reaps the benefit of low computation complexity and high accuracy of expert joint decision. Experiments on two benchmark datasets confirmed the effectiveness of the proposed method.

5. ACKNOWLEDGMENTS

This work is supported by the State Key Program of National Natural Science of China (Grant No. 61232010), the National Natural Science Foundation of China (Grant No. 61172143, 61379094 and 61105012), and the Fundamental Research Funds for the Central Universities (Grant No. 3102014JC02020G07) .

Table 1: Quantitative comparison of the proposed method with the state-of-the-arts for abnormal detection on the UMN dataset.

Method	AUC
Optical Flow [9]	0.8400
Social Force Model [9]	0.9600
Streakline Potentials[15]	0.9000
Sparse-Lawn [4]	0.9955
Sparse-Indoor [4]	0.9750
Sparse-Square [4]	0.9640
Ours-Lawn	**0.9928**
Ours-Indoor	**0.9700**
Ours-Square	**0.9879**
Ours-All	**0.9764**

6. REFERENCES

[1] PETS2009 dataset. `http://www.cvg.rdg.ac.uk/PETS2009/a.html`.

[2] UMN dataset. `http://mha.cs.umn.edu/movies/`.

[3] H. Cheng and J. Hwang. Integrated video object tracking with applications in trajectory-based event detection. *J. Visual Communication and Image Representation*, 22(7):673–685, 2011.

[4] Y. Cong, J. Yuan, and J. Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 2012.

[5] X. Cui, Q. Liu, M. Gao, and D. Metaxas. Abnormal detection using interaction energy potentials. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 3161–3167, Jun. 2011.

[6] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *Proc. British Machine Vision Conference*, pages 1–11, Sept. 2010.

[7] O. Junior, D. Delgado, V. Gonçalves, and U. Nunes. Trainable classifier-fusion schemes: an application to pedestrian detection. In *Proc. IEEE Conf. Intelligent Transportation Systems*, pages 1–6, Oct. 2009.

[8] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1975–1981, Jun. 2010.

[9] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 935–942, Jun. 2009.

[10] Q. Wang, J. Fang, and Y. Yuan. Multi-cue based tracking. *Neurocomputing*, 131:227–236, 2014.

[11] Q. Wang, Y. Yuan, P. Yan, and X. Li. Saliency detection by multiple-instance learning. *IEEE T. Cybernetics*, 43(2):660–672, 2013.

[12] S. Wang and Z. Miao. Anomaly detection in crowd scene. In *Proc. Int. Conf. Signal Processing*, pages 1220–1223, Oct. 2010.

[13] S. Wu, B. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2054–2060, Jun. 2010.

[14] Y. Yuan, J. Fang, and Q. Wang. Robust superpixel tracking via depth fusion. *IEEE Trans. Circuits Syst. Video Techn.*, 24(1):15–26, 2014.

[15] Y. Zhang, L. Qin, H. Yao, and Q. Huang. Abnormal crowd behavior detection based on social attribute-aware force model. In *Proc. Int. Conf. Image Processing*, pages 2689–2692, Oct. 2012.