Unsupervised Fast Anomaly Detection in Crowds

Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, Xianming Liu, Pengfei Xu
Department of Computer Science, Harbin Institute of Technology
No.92, West Dazhi Street, Harbin, P. R. China, 150001
{xiaoshuaisun, h.yao, rrji, xmliu, pfxu}@hit.edu.cn Tel: +86-451-86416485

ABSTRACT
In this paper, we proposed a fast and robust unsupervised framework for anomaly detection and localization in crowed scenes. Our method avoids modeling the normal state of the crowds which is a very complex task due to the large within class variance of the normal target appearance and motion patterns. For each video frame, we extract the spatial temporal features of 3D blocks and generate the saliency map using a block-based center-surround difference operator. Then, motion vector matrix is obtained by adaptive rood pattern search block-matching algorithm and distance normalization. Attractive motion disorder descriptor is proposed to measure the global intensity of anomalies in the scene. Finally, we classify the frames into normal and anomalous ones by a binary classifier. In the experiments, we compared our method against several state-of-the-art approaches on UCSD dataset which is a widely used anomaly detection and localization benchmark. As the only unsupervised approach, our method outputs competitive results with near real-time processing speed.

Categories and Subject Descriptors
H.3.1 [**Information Systems**]: Content Analysis and Indexing; H.5.1 [**Multimedia information Systems**]: Video

General Terms
Algorithms, Experimentation, Human Factors

Keywords
Unsupervised anomaly detection, motion estimation, attractive motion disorder descriptor

1. INTRODUCTION
As reviewed in [1, 2], monitoring surveillance videos, especially for videos of crowded scene, is a very expensive and tiring task. Thus, automatic detection of anomalous events in crowds has become an attractive topic in computer vision and pattern recognition research. Due to the unreliability of trajectory analysis in crowded scene [3], recent works focus on designing robust dynamic scene representations that avoid multiple targets tracking [4, 5, 6, 7, 8]. Adam *et al.* [4] maintain probabilities of optical flow in local regions, using histograms. Kim and Grauman

*Area Chair: Kiyoharu Aizawa

MM'11, November 28-December 1, 2011, Scottsdale, Arizona, USA.

(a) Normal moving targets (b) The crowed scene

Figure 1. Large within-class variance of the normal target appearance and motion patterns in crowded scene.

[5] utilized a mixture of probabilistic **PCA** models to model local optical flow patterns, and enforce global consistency using a Markov Random Field (**MRF**). Inspired by classical studies of crowd behavior, Mehran *et al.* [6] characterized crowd behavior using concepts such as social force. These concepts lead to optic flow measurement of target interaction within the crowds, which are combined with a latent Dirichlet Allocation (**LDA**) model for anomaly detection. Mahadevan *et al.* [8] proposed a unified framework for joint modeling of appearance and dynamics of the scene, under which the outliers are labeled as anomalies.

However, scene representation is not the only problem for anomaly detection task. Modeling the normal state of the crowded scene is another challenging problem due to the large within-class variance of the normal target appearance and motion patterns. Figure 1 shows the moving targets appeared in a 20 seconds video clip, which contains different target appearances and movements. In real-world applications, the length of the video with normal crowd behaviors will be much longer than 20 seconds, thus it's nearly impossible to model the normal state containing thousands of patterns with different spatial temporal appearance. Compared with supervised or semi-supervised learning of the normal states [2, 3, 4, 5, 6, 7, 8, 9], it may be more practical to directly model the global intensity of anomalous events in a purely unsupervised manner.

From experimental observations, we found that abnormal contents or unusual human behaviors will consistently attract the attention of human observers, which means most of the anomalies are more attractive or more salient compared with the other contents in the environment. Besides, the presence of anomalies will probably turn the ordered crowd movements into a disordered state. Based on these observations, we proposed an unsupervised framework for anomaly detection and localization task, which uses *Attractive Motion Disorder descriptor* to directly measure the overall intensity of anomalies and avoids modeling of the crowd's normal behavior. Our descriptor is constructed by fusing the statistical features of visual saliency and motion vectors, which is inspired by both the perceptual and computational observations on normal and anomalous videos.

2. METHOD

The proposed unsupervised anomaly detection framework is shown in Figure 2. Temporal derivatives of spatial temporal video blocks are extracted as visual features. Saliency is then computed by block-based center-surround difference operator. Motion disorder is measured by the standard deviation of the motion vectors estimated by adaptive block-match algorithm. By analyzing the statistical distribution of visual saliency and the motion vector matrix, we construct attractive motion disorder descriptor to measure the global anomalous intensity, with which video frames are classified into normal or anomalous frames by a binary classifier. Localization of the detected anomalies is achieved using the spatial temporal saliency map.

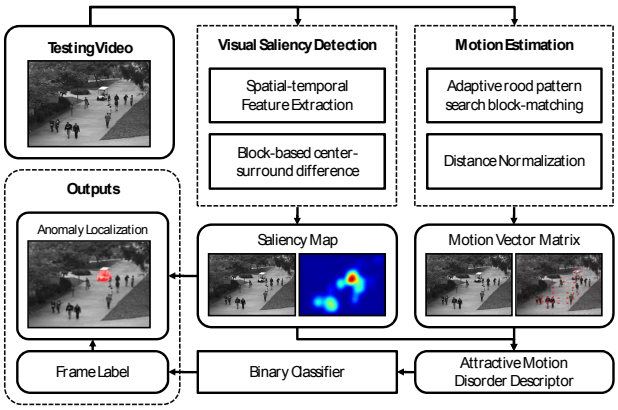

Figure 2. The framework of the proposed method.

2.1 Center-Surround Saliency Detection

Saliency is an important concept for computational visual attention modeling, which could be quantitatively measured by center-surround difference [10, 11], information maximization [12], incremental coding length [13] and site entropy rate [14], *etc*. In our case, we first extract spatial-temporal local features from the video, then generate the saliency map using a block-based center-surround operation, which is more computational efficient and shares the plausibility of previous works. The visual field is segmented into 24×32 3D sub-blocks represented by a gradient-based spatial-temporal descriptor. The descriptor of a sub-block is constructed by the absolute values of the temporal derivatives in all pixels in the block. These values are stacked into a 1-D feature vector. A center-surround difference operator, akin to the visual receptive fields of human vision system, is adopted as a quantitative measurement for visual saliency. In traditional models [10, 11], the center-surround difference was computed across different spatial scales using Difference of Gaussian filters. In our case, we only compute the difference between center block and its surrounding eight-neighborhoods for the concern of computation efficiency. The saliency of a given block is defined as the average center-surround difference measured by the Manhattan Distance between the features of the center and its surrounding blocks:

$$S_{i,j} = \sum_{m=-1}^{1} \sum_{n=-1}^{1} \left| F_{i,j} - F_{i+m,j+n} \right|_1 \qquad (1)$$

Figure 3 shows some examples of the spatial temporal saliency maps computed using the temporal gradient features and block-based center-surround difference operator. It's easy to notice that the anomalies tend to appear at the locations with the largest saliency value in the scene.

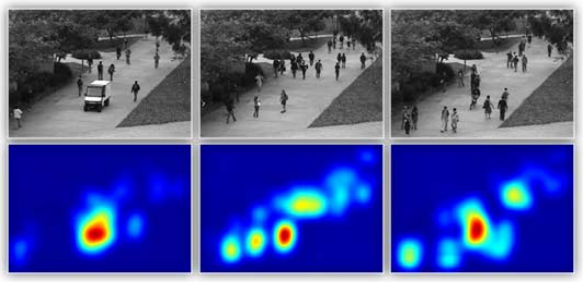

Figure 3. Examples of the spatial temporal saliency maps.

2.2 Attractive Motion Disorder Descriptor

The motion vectors obtained by adaptive rood pattern search block-matching algorithm [15] are used as motion descriptors for each sub-block. The visual field is segmented into 12×16 sub-blocks with equal size. Note that, motion vectors can also be directly obtained from the compression domain data if the video is compressed using motion compensation technique. Let $M_{i,j}$ denote the motion vector of the sub-block in *i*th row and *j*th column, we apply distance normalization to eliminate the scale variance of the motion vector caused by the geometrical setting of the camera:

$$M'_{i,j} = \left(\frac{\beta \cdot (H - i)}{H} + \beta \right) \cdot M_{i,j}, \qquad (2)$$

where M is the motion vector matrix, H is the height of M, $\beta = 0.5$ is a distance compensation parameter which has been fixed in our experiment. After normalization, object moments appeared in all sub-blocks can be near equally measured by *M'*. Figure 4 illustrates motion estimation and normalization results.

Figure 4. Motion estimation. From left to right: input video frame, motion estimation result by adaptive rood pattern search block-matching [15] and normalized motion vectors.

There are various measurements for system disorder such as *Entropy* and *Standard Deviation*. *Entropy* is an important concept in physics and information theory, which is also widely used as a quantitative measurement for uncertainty or unpredictability. *Standard Deviation* is an easy to compute statistical feature describing the variance or diversity of a group of data. Practically, we use standard deviation to measure the motion disorder, because it leads to a better overall performance while costing much less computations compared with other measurements. Given the spatial temporal saliency map *S*, the motion vector matrix *M',* we define the **A**ttractive **M**otion **D**isorder (**AMD**) descriptor *A* by:

$$A = \alpha \cdot \max(S) + (1 - \alpha) \cdot \mathrm{std}(M'), \qquad (3)$$

where $\alpha \in [0,1]$ is a fusing parameter, std(.) denotes the standard deviation of the input matrix. The descriptor can be regarded as a quantitative measurement for global intensity of all the anomalous events appeared in the visual field. Higher value for the **AMD** descriptor indicates larger probability for the appearance of anomalies. Figure 5 illustrate the distribution of **AMD** descriptor ($\alpha \in 0.5$) in normal and anomalous videos.

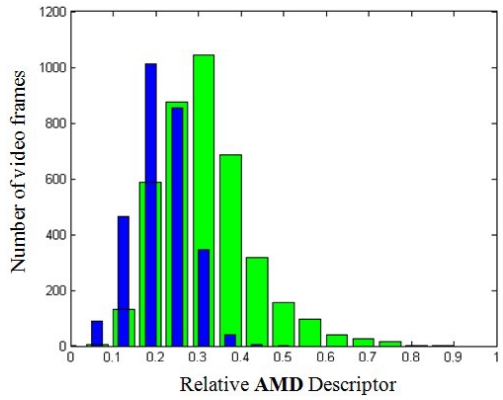

Figure 5. Distribution of AMD descriptor in normal (blue) and anomalous (green) video frames of UCSD Ped_1 dataset.

2.3 Anomaly Detection and Localization

Video frames can be classified into normal or abnormal frames by a binary classifier using the **AMD** descriptor. As described in Section 1, anomalous regions tend to attract more visual attention compared with the other events happened in the scene. Thus, saliency map can be used as a reference for localization and segmentation of the anomalous regions. In practice, we adopt Equation 4 to segment the anomalies, which is first proposed in [16] for non-parametric proto-object segmentation.

$$\mathrm{O}(x,y)\begin{cases}1 & \text{if } S'(x,y) > \text{threshold},\\ 0 & \text{otherwise.}\end{cases} \tag{4}$$

where O is the localization binary map, $S' = S * G$ is a refined saliency map smoothed by a Gaussian filter G ($3\times3, \sigma = 1$). We set the threshold to be $7\times\mathbf{E}(S)$ empirically, where $\mathbf{E}(S)$ is the mean intensity of the saliency map. Examples of anomaly detection and localization results are shown in Figure 6.

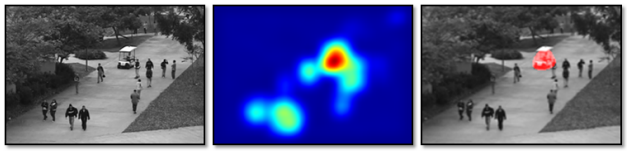

Figure 6. Anomaly detection in crowded videos. From left to right: Detected abnormal frame, corresponding saliency map and localization result of the anomalous region.

3. EXPERIMENTS

We evaluate the proposed approach on UCSD dataset [8][1], which is a well annotated publicly available dataset for the evaluation of abnormal detection and localization in crowded scenes. The dataset was acquired with a stationary camera mounted at an elevation at a resolution of 238 × 158 with 10 fps, overlooking pedestrian walkways. The circulation of non pedestrian entities in the walkways, and anomalous pedestrian motion patterns are regarded as abnormal events. Commonly appeared anomalies include bikers, skaters, small carts, and people walking across a walkway or in the grass. Videos were split into 2 subsets: **Ped_1** and **Ped_2**, each corresponding to a different scene. Videos recorded from each scene were split into various clips each of which has around 200 frames. **Ped_1** contains 34 training clips and 36 testing clips, while **Ped2** contains 16 training clips and 14 testing clips. For each clip, the ground truth annotation includes a binary flag per frame, indicating whether an anomaly is present in that frame.

Practically, all video frames are resized to 120×160 in order to reduce the computation cost. For each frame, we extract the spatial temporal features of 5×5×3 3D video blocks, and generate a 24×32 saliency map using the proposed block-based center-surround difference operator. A 12×16 motion vector matrix is then obtained based on adaptive rood pattern search block-matching algorithm and distance normalization. Based on the saliency map and the motion vector matrix, we compute the **AMD** descriptor using Equation 3 ($\alpha = 0.5$), which is proposed to describe the overall intensity of anomalies appeared in the frame. Finally, the video frame is classified into normal or anomalous frame by a binary classifier.

The evaluation on **UCSD** dataset contains two components: anomaly detection and localization. By varying the parameters of the tested approach, an **ROC** curve can be drawn to intuitively evaluate the anomaly detection performance. Figure 7 illustrates the **ROC** curves for **UCSD** dataset of various state-of-the-art approaches and our approach, while Figure 8 shows some visual examples of anomaly localization and segmentation results of the tested approaches. In addition to Figure 7, Table 1 shows the area under **ROC** curve (**AUC**) of the tested methods, in which a larger **AUC** score means better classification performance. According to the experimental results, our method, as the only completely unsupervised training-free approach, outputs competitive results against the state-of-the-art methods with near real-time processing speed. Visual results indicate that our method is able to accurately localize the anomalous events in the crowded scene and outputs better segmentation results with well defined boundaries.

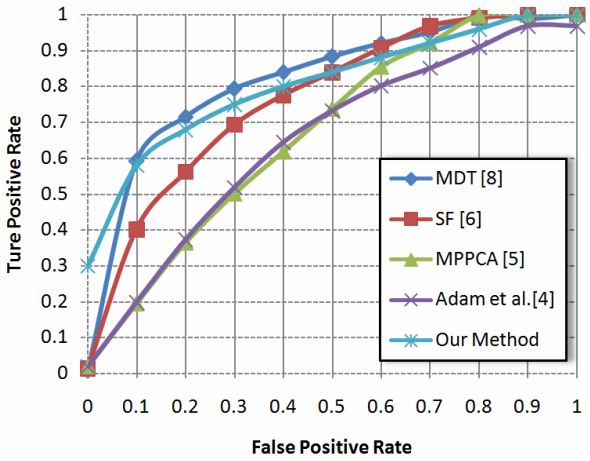

Figure 7. ROC curves of tested approaches on UCSD Ped_1 dataset. Tested approaches include our method, MDT-based approach [8], the Social Force Model [6], the mixture of optical flow (denoted as MPPCA [5]) and optical flow monitoring method (Adam *et al.* [4]).

Table 1. Area Under ROC Curves

Method	MDT	SF	MPPCA	Adam	Ours
AUC	0.7895	0.7413	0.6554	0.6350	**0.7919**

[1] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

4. CONCLUSION

In this paper, we proposed an unsupervised framework for fast anomaly detection and localization in crowded scene. Instead of modeling the normal states, we directly model the intensity of anomalies using attractive motion disorder descriptor, which is constructed by fusing the statistical features of saliency map and motion vector matrix. Saliency detection and motion estimation are conducted by block-based center-surround difference operator and adaptive rood pattern search block-matching algorithm, both of which are highly efficient and lead to a near real-time overall processing speed. Experimental results on a widely used benchmark dataset demonstrate the effectiveness of the proposed framework. Our future work lies in integrating other reliable features, such as location distribution prior, into the framework to further improve the overall performance.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61071180 and Key Program Grant No. 61133003).

6. REFERENCES

[1] N. Haering, P. Venetianer, and A. Lipton. "The evolution of video surveillance: an overview". *Machine Vision and Applications*, 19(5-6):279–290, 2008.

[2] L. Seidenari, M. Bertini. "Non-parametric anomaly detection exploiting space-time features". *ACM Multimedia*, pp.1139–1142, 2010.

[3] F. Jiang, Y. Wu, and A. Katsaggelos. "A dynamic hierarchical clustering method for trajectory-based unusual video event detection". *IEEE TIP*, 18(4):907–913, 2009.

[4] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. "Robust real-time unusual event detection using multiple fixed location monitors". *IEEE TPAMI*, 30(3):555–560, 2008.

[5] J. Kim and K. Grauman. "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates". *CVPR*, pp. 2921–2928, 2009.

[6] R. Mehran, A. Oyama, and M. Shah. "Abnormal crowd behavior detection using social force model". *CVPR*, pp.935–942, 2009.

[7] L. Kratz and K. Nishino. "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models". *CVPR*, pp.1446–1453, 2009.

[8] V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos. "Anomaly Detection in Crowded Scenes". *CVPR*, 2010.

[9] O. Boiman and M. Irani. "Detecting irregularities in images and in video". *IJCV*, 74(1):17–31, Aug. 2007.

[10] L. Itti, C. Koch and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis". *IEEE TPAMI*, 20(11), 1998.

[11] D. Gao, V. Mahadevan, and N. Vasconcelos. "The discriminate center-surround hypothesis for bottom-up saliency". *Advances in Neural Information Processing Systems*, pp.497-504, 2007.

[12] N. Bruce and J. Tsotsos. "Saliency based on information maximization". *Advances in Neural Information Processing Systems*, pp.155–162, 2006.

[13] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments. *NIPS,* pp. 681–688, 2008.

[14] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring Visual Saliency by Site Entropy Rate". *CVPR*, pp. 2368–2375, 2010.

[15] Y. Nie, and K. Ma. "Adaptive rood pattern search algorithm for fast block matching motion estimation". *IEEE TIP*. 11(12), pp.1442--1448, 2002.

[16] X. Hou and L. Zhang, Saliency detection: a spectral residual approach. *CVPR*, 2007.

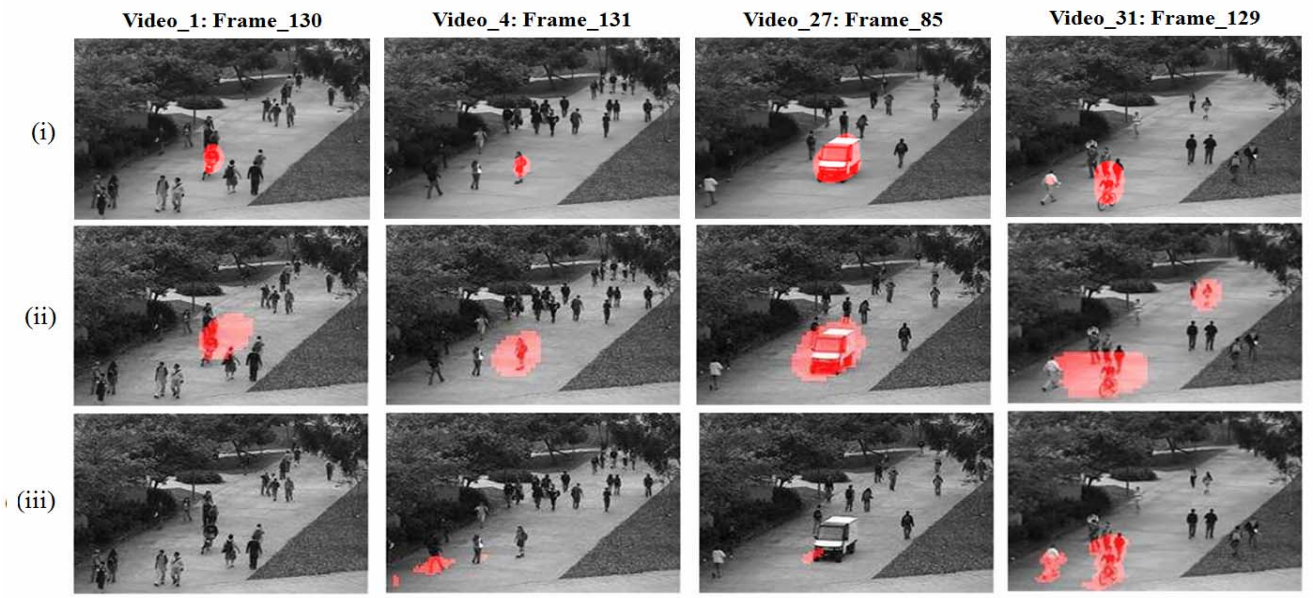

Figure 8. Comparisons of abnormal localization results from (i) our approach; (ii) MDT approach and (iii) SF-MPPCA approach. The results of MDT and SF-MPPCA are provided by Mahadevan *et al.* [8]