HIERARCHICAL ACTIVITY DISCOVERY WITHIN SPATIO-TEMPORAL CONTEXT FOR VIDEO ANOMALY DETECTION

Dan Xu[1], *Xinyu Wu*[1,3], *Dezhen Song*[2], *Nannan Li*[1], *Yen-Lun Chen*[1]

[1]Guangdong Provincial Key Lab of Robotics and Intelligent System,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2]Department of Computer Science and Engineering, Texas A&M University, USA
[3]Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong

ABSTRACT

In this paper, we present a novel approach for video anomaly detection in crowded and complicated scenes. The proposed approach detects anomalies based on a hierarchical activity pattern discovery framework comprehensively considering both global and local spatio-temporal contexts. The discovery is a coarse-to-fine learning process with unsupervised ways for automatically constructing normal activity patterns at different levels. An unified anomaly energy function is designed based on these discovered activity patterns to identify the abnormal level of an input motion pattern. We demonstrate the efficiency of the proposed method on the UCSD anomaly detection datasets (Ped1 and Ped2) and compare the performance with existing work.

Index Terms— Visual surveillance, video anomaly detection, hierarchical discovery, energy function

1. INTRODUCTION

Video anomaly detection has become an important research aspect in the area of intelligent visual surveillance due to the growing security needs. Many researchers have been focusing on this area in recent years. However, the uncertainty of the abnormal activity description and the complexity of the scenes make the anomaly detection a challenging problem.

For detecting anomaly, one category of popular approaches in the literature is the tracking-based methods [1] [2] [3]. The main idea of such methods is to analyze and model normal trajectories collected by tracking individual moving objects in the video, then to detect anomalous object motions whose trajectories are deviating from the normal model. These methods can obtain promising results under the less cluttered scenes with only a few people, but in dense crowds, to achieve robust tracking is a quite difficult task because of serious occlusion problems, which heavily degrades the performance of the anomaly detection.

To avoid the aforementioned limitations, the other category of methods address the problem by learning activity patterns from low-level visual features. Andrade et al. model crowd scenes with Hidden Markov Models combined with spectral clustering for detecting unusual events [4]. Mehran et al. propose to model the crowd activity patterns for the anomaly detection using a “social force” model based on optical flow feature representation [5]. Kim et al. use a mixture dynamic texture model to detect spatial and temporal anomalies through a joint modeling of appearance and dynamics of the scene [6]. However, these methods model activity patterns only considering the local context or global context, which leads to the lack of global information or local location relationship for the simultaneous perception of both local and global abnormal motion pattern.

In this paper, we aim to detect anomalies comprehensively considering both global and local spatio-temporal contexts. A hierarchical framework of learning activity pattern is proposed to achieve this task. Under the global context, we discover atomic activity patterns from low-level optical flow features, and the distributions of the atomic activity patterns are modeled for higher-level activity representation. Then salient activity patterns are discovered under the local context. The two layers of discovery both adopt unsupervised ways without any priori knowledge of the anomaly. Finally, we design an unified abnormal energy function to detect global and local pattern anomalies. The overview of our proposed approach is illustrated in Fig. 1.

2. HIERARCHICAL ACTIVITY DISCOVERY

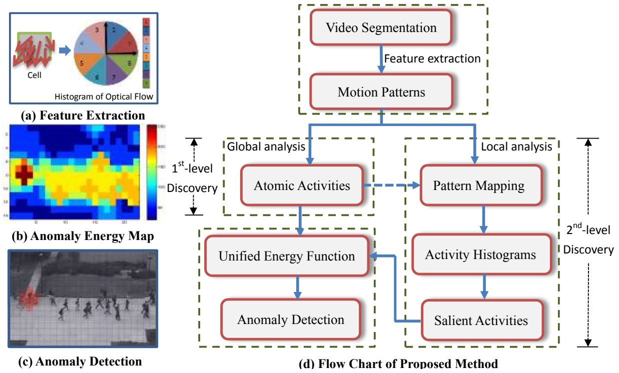
2.1. Feature Representation

For anomaly detection, we first extract features in the video. For a video frame with a resolution of $w \times h$, it is divided into non-overlapping cells with a size of $L \times L$. In the paper, an effective motion feature represented by the direction and magnitude of the optical flow is used as a low-level feature for the underlying motion pattern description. The optical flow field is calculated at every pixel position using the algorithm proposed by Liu et al. [7]. To obtain a joint representation of motion direction and speed, for any cell $c\{i,j\}(i \in \{1,...,w/L\}, j \in \{1,...,h/L\})$, a 8-dimension motion feature vector is extracted through the accumulation of the optical flow magnitude of every pixel corresponding to 8 different direction intervals in this cell, as shown in Fig. 1 (a).

2.2. Discovery of atomic activities in global context

For all training video frames, cell-based motion feature vectors are extracted from every cell position with the method presented in Subsection 2.1. These feature vectors represent low-level motion pat-

The work described in this paper is partially supported by the National Natural Science Foundation of China (61005012), Shenzhen Fundamental Research Program (JC201105190948A), and Guangdong Innovative Research Team Program (201001D0104648280).



(a) Feature Extraction

(b) Anomaly Energy Map

(c) Anomaly Detection

(d) Flow Chart of Proposed Method

Fig. 1: The overview of our approach for anomaly detection.

terns, which are used to discover atomic activity patterns globally using unsupervised learning. The atomic activity is a basic unit of activity patterns. We choose the K-means algorithm to group all normalized feature vectors into k clusters with centers $\{g_1, ..., g_k\}$. These clusters stand for global and normal atomic activities which occur within the training video frequently.

To model each atomic activity, the non-parametric kernel density estimation is used to generate smooth probability density function. Due to the high dimensional curse, we estimate the distribution of each cluster by using distances of points in the cluster to the center instead of 8-dimensional feature vectors. Given a collection of feature vectors $\{I_1, ..., I_n\}$ from cluster i, let $x_j = E(I_j, g_i)$ denote the Euclidean distance between a feature vector I_j and the center g_i, then the probability of a newly observed value x can be estimated using a Parzen window density estimation approach with a Gaussian kernel [8] as follows:

$$\widehat{P_d}(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h\sqrt{2\pi}}exp\{-\frac{1}{2}(\frac{x-x_j}{h})^2\}, \tag{1}$$

where $h > 0$ is the bandwidth corresponding to the Gaussian kernel.

2.3. Discovery of salient activities in local context

Since activity patterns are highly related to the location, a normal activity pattern in an area can be anomalous in another area. For example, a people walking on the sideway is a normal activity while walking on the lawn can be deem as anomalous. So we propose to discover salient activities considering from local spatio-temporal context.

For a motion feature I, a procedure of pattern mapping is performed. We calculate a membership degree vector $s(I)$ with the k probability distributions $\{\widehat{P_{d1}}, ..., \widehat{P_{dk}}\}$, $s(I) = [\widehat{P_{d1}}(E(I, g_1)), ..., \widehat{P_{dk}}(E(I, g_k))]$. $s(I)$ can be considered as a higher-level activity representation upon global atomic activity patterns and we call it an activity histogram. For a cell $c\{i, j\}$, take into account the high similarity of activity patterns between adjacent cells, a set of activity histograms $S_{i,j} = \{s_1, ..., s_{5N}\}$ is obtained from not only itself but also its four neighbor cells within all N training frames. The goal is to learn M salient activity patterns $H_{i,j} = \{h_1, ..., h_M\}$ from $S_{i,j}$. $H_{i,j}$ need to represent $S_{i,j}$ sparsely ($M \ll 5N$) and compactly (with the maximum similarity to the original set). Moreover, since the number of salient activity patterns at various cell positions may distinctly vary from each other, for example, in a sky region and in a sidewalk region. So it needs to determine the value of M adaptively during learning salient activity patterns at each cell position.

To satisfy these requirements, the hierarchical clustering technique is used as the basic framework for handling this local discovery task. For hierarchical clustering, the choice of dissimilarity metric is the key aspect for the learning performance. To consider the spatial correlation of the global atomic activities, we introduce the Earth Mover's Distance (EMD) for the distance measure of two activity histograms. The EMD is an efficient cross-bin dissimilarity measure method with high robustness for the histogram comparison, which has been successfully applied in image retrieval [9] and visual tracking [10]. For two activity histograms p and q which are normalized to unit mass, the EMD between them can be obtained through solving the transportation problem: $\min \sum_{i,j=1}^{k} f_{ij}d_{ij}$, subject to $f_{ij} \geq 0, \sum_{j=1}^{k} f_{ij} = p^i$, and $\sum_{i=1}^{k} f_{ij} = q^j$, where f_{ij} denotes the flow from i-th supply in p to j-th demand in q and $d_{i,j}$ denotes the ground distance. To reduce the time complexity of the original EMD, we apply the algorithm proposed in [9] for fast calculation (from $O(N^3 \log N)$ to $O(N)$). The details of the discovery of local salient activity patterns are illustrated in Algorithm 1.

Algorithm 1 Discovery of local salient activity patterns

```
Input: $S_{i,j} = \{s_1, ..., s_{5N}\}$ for a cell$\{i, j\}$
Output: $H_{i,j} = \{h_1, ..., h_M\}$
1: Construct agglomerative hierarchical clustering tree $A_{tree}$ from
   $S_{i,j}$ with EMD distance metric;
2: Determine the number of clusters, $M$. An evaluation graph (#
   of clusters vs. EMD evaluation metric) is produced with $A_{tree}$.
   Let $b = 5N$, $L_m$ denote the points with $x = 2, ..., m$ and
   $R_m$ the points with $x = m+1, ..., b$, $M = \arg\min_{m}\{\frac{m-1}{b-1} \times$
   $\text{RMSE}(L_m) + \frac{b-m}{b-1} \times \text{RMSE}(R_m)\}$, where RMSE is the root
   mean squared error of the best fit-line for the sequences of points
   in $L_m$ or $R_m$ [11];
3: Discover a salient activity $h_i (i = 1, ..., M)$ for each cluster.
   Let $x_k$ denote $k^{th}$ dimensional value of points in cluster $i$, and
   $[h_i]_k$ denote $k^{th}$ dimension of $h_i$. $[h_i]_k = \mu_k$ by $(\mu_k, \sigma_k) =$
   $\arg\max_{\mu_k,\sigma_k} G(x_k|\mu_k, \sigma_k)$, where $G(x)$ is a Gaussian distribution;
4: Return: A set of local salient activity patterns $\{h_1, ..., h_M\}$
```

3. UNIFIED ENERGY FUNCTION FOR ANOMALY DETECTION

The anomaly detection in the paper is a cell-based binary classification problem. The possible anomaly cases includes: global pattern anomaly (i.e. a motion pattern with low similarity to the global atomic activity patterns), local pattern anomaly and the co-occurrence of global and local pattern anomaly. We decide whether an anomaly occurs within a cell $c\{i, j\}$ through a combination of two-layer dissimilarity measure evaluated from the test motion pattern compared with the discovered global and local activity patterns, respectively.

To detect global pattern anomaly in cell$\{i, j\}$, we first calculate the posterior probability of each global atomic activity pattern g_i for a given motion vector I with the Bayes' rule as follows: $P(g_i|I) = \frac{P(I|g_i)P(g_i)}{\sum_{n=1}^{k} P(I|g_n)P(g_n)}$, $i = 1, 2, ..., k$, where the priori probability of

$P(g_i)$ is represented by the ratio of the number of the samples in cluster i to the number of all training samples, and the likelihood $P(I|g_i)$ is calculated with Equation (1) by $P(I|g_i) = \widehat{P_d}(E(I, g_i))$. Based on these, we can obtain a global atomic activity pattern i' which produces the maximum probability for the motion vector I, i.e. $i' = \arg\max_i P(g_i|I)$. Then the level of global pattern anomaly for the cell can be given by

$$Ag(I) = \frac{1}{\log P(g_{i'}|I)}. \tag{2}$$

For local pattern anomaly, let $Al(I)$ denote the dissimilarity of the activity histogram $s(I)$ with the local salient activity patterns $\{h_1, ..., h_M\}$ corresponding to cell $c\{i, j\}$, and α_m denote the ratio of the number of samples in the cluster m to the number of samples in the training set $S_{i,j}$. Given a weight vector $\{\alpha_1, ..., \alpha_M\}$, Al is calculated with Equation (3).

$$Al(I) = \sum_{m=1}^{M} \alpha_m \text{EMD}\{s(I), h_m\}, \tag{3}$$

where EMD$\{a, b\}$ denotes the earth mover's distance measure between histogram a and histogram b. Then Al can be treated as the level of local pattern anomaly.

Finally, we design an unified energy function for the anomaly detection task. Since the bigger the value Ag is, the higher probability the global pattern anomaly occurs with, while the bigger value Al is, the higher probability local pattern anomaly occurs with. Hence, an estimation of abnormal energy for a cell can be given with the form of the product by $Ag * Al$. To consider the spatial-temporal correlation of the motion pattern, we measure the final abnormal energy of the cell by an average estimation within a 3D spatial-temporal volume which contains the cell and its adjacent four cells as well as the same regions of the previews and the next T frames. Then the abnormal energy function are designed as follows:

$$A(I) = \frac{\sum_{t=1}^{2T+1} \prod_{n=1}^{5} \{Ag_t^n(I) * Al_t^n(I)\}}{2T + 1}. \tag{4}$$

Here, the value of T is set to 1 in our algorithm. Then for a given anomaly energy threshold A_θ, if $A(I) > A_\theta$, this cell is classified as an abnormal region.

4. PERFORMANCE EVALUATION

To evaluate the performance of the proposed approach, we carry out anomaly detection experiments with the public available UCSD anomaly detection dataset [12]. This dataset consists of two subsets (Ped1 and Ped2), which are captured under two different scenes with anomalies including bikers, vehicles, skateboarders and people walking on the lawn. The anomalies occur at multiple different locations in some of frames. The image sizes of Ped1 and Ped2 are 238×158 and 360×240, respectively. Ped1 has 34 image sequences for training and 16 for testing and Ped2 has 16 image sequences for training and 12 for testing. The test set in Ped1 has about 3400 anomalous frames and 5500 normal frames, while in Ped2 about 1652 anomalous frames and 346 normal frames.

According to the image resolutions, the cell size for Ped1 and Ped2 is set to 10 × 10 and 15 × 15, respectively. Then the grid matrix sizes of images for them are both 24 × 16. For discovering the global atomic activity patterns, the number of patterns k is set to 40. The proposed algorithm is implemented with Matlab. The code of optical flow calculation is written in C++ and mexed to be called in Matlab for improving the computational efficiency [7]. We test the algorithm on Ped1 and Ped2 datasets, and Figure 3 shows examples of our anomaly detection results.

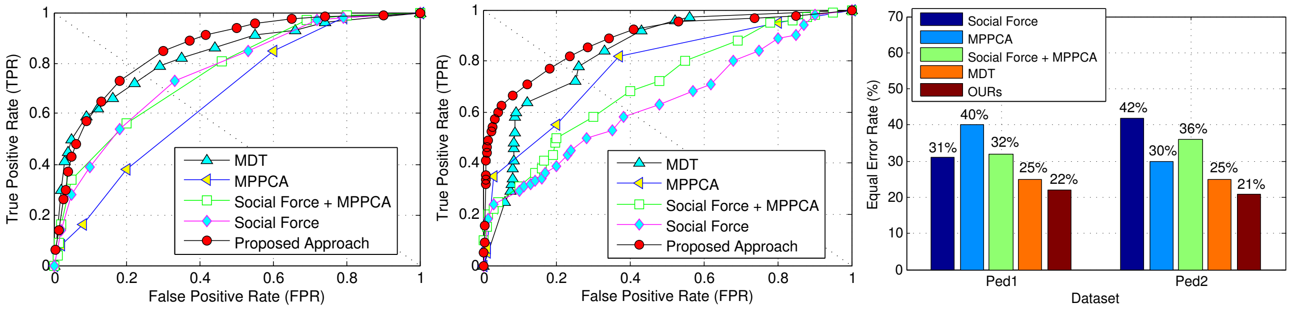
Algorithm	Social Force [5]	MPPCA [6]	Social Force + MPPCA [12]	MDT [12]	OURs
Ped1	67.5%	59.0%	66.8%	81.8%	**85.4%**
Ped2	55.6%	69.3%	61.3%	82.9%	**88.2%**
Average	61.6%	64.2%	64.1%	82.4%	**86.8%**

Table 1: The quantitative comparison of the AUC (Area Under ROC) on Ped1 and Ped2.

Quantitative evaluation results. Given an abnormal energy map of each test frame, the anomaly regions are detected via a predefined threshold value. We conduct anomaly detection experiments with different threshold values based on a frame-level groundtruth. For each testing image, the groundtruth annotation uses a binary flag to represent whether one or more anomalies are present. Then two ROC curves are produced over Ped1 and Ped2 datasets, as shown in Fig. 2 (a) and (b). Fig. 2 (c) shows the Equal Error Rate (EER) of our algorithm. For performance comparison, we choose four state-of-the-art methods including the Mixture Dynamic Texture (MDT) [12], the Mixture of Optical Flow (MPPCA) [6], the social force [5] and the social force with MPPCA [12]. The quantitative results of these four methods are obtained from the paper [12]. From Fig. 2, we can note that our algorithm outperforms all other four methods at the equal error rate on both Ped1 and Ped2 datasets. We also calculate the Area Under Curve (AUC) (i.e. the area under ROC) shown in Table 1. The average AUC of our algorithm on the two test datasets is 86.8%, which is 4.4% higher than 82.4% of the MDT, the best one of four approaches for comparison.

Computational efficiency. The speed performance of the algorithm is an important aspect for practical application consideration. Under a standard PC platform with 3 GHz CPU and 2 GB memory, the processing time of our algorithm is 5 seconds per frame under the Matlab environment (i.e 12 frames per minute), while the processing time of the MDT approach is 25 seconds per frame (i.e 2.4 frames per minute). In the proposed algorithm, since the main time and storage overhead is from the offline phase for the hierarchical discoveries of activity patterns, so we can perform faster during the online test phase. We are also able to improve the test speed further by only calculating and analyzing the moving pixels in the test frame with the dynamic background substraction scheme.

Analysis. The experiments demonstrate the anomaly detection performance of the proposed algorithm. Our approach can achieve not only high frame-level anomaly detection rate, but also accurate anomaly localization as illustrated in Fig. 3. That is because for each cell location within spatio-temporal context, the algorithm discovers normal activity patterns from different levels, which are further combined for detecting anomalous activity patterns. In addition, from the low-level global atomic activity (8-dimensional vector) to the higher-level salient activity pattern (40-dimensional vector), it is a dimension raising process that can help to improve the discrimination for anomalous patterns and realize fine-grained anomaly detection at lo-



(a) ROC curve of PED1 Dataset

(b) ROC curve of PED2 Dataset

(c) Equal Error Rate of PED1 and PED2

Fig. 2: The quantitative comparison of anomaly detection results on Ped1 and Ped2 datasets.

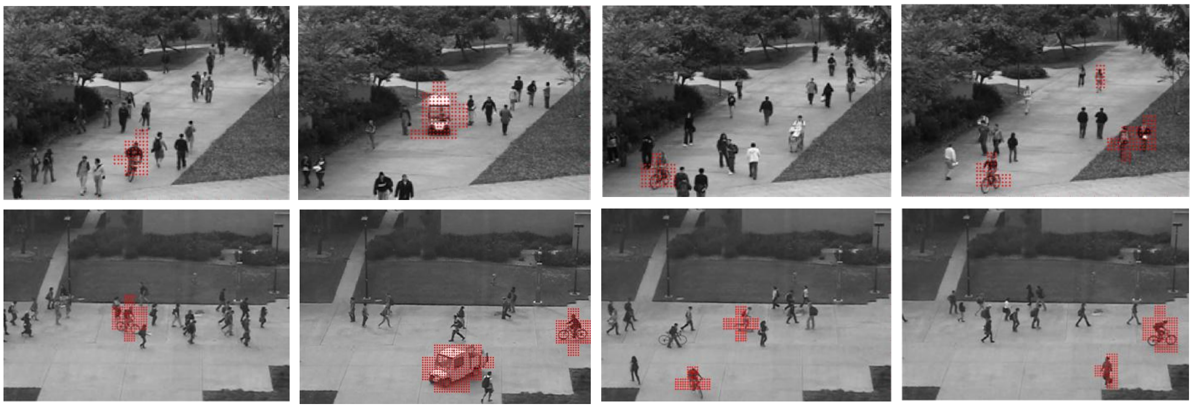


Fig. 3: Examples of anomaly detection results with the proposed approach on Ped1 (the first row) and Ped2 (the second row) datasets. The abnormal regions (cells) are marked with red dots. The proposed approach can detect anomalies such as bikes, vehicles, skaters and people walking on the lawn with accurate localization of the anomalies.

cal locations. However, there are also 'miss' (True Negative) cases occur in our detection results. For example, as shown in Fig. 3, a man pushing a shopping cart and a man walking with a bicycle slowly are classified as normal activities. Because our algorithm is based on motion feature described with optical flow, which cannot distinguish them from the normal pedestrian activity patterns.

5. CONCLUSION

In this paper, we have presented a video anomaly detection approach suitable for crowded and complicated scenes. A hierarchical framework of activity pattern discovery was proposed with the comprehensive consideration of global and local spatio-temporal contexts, which can learn normal atomic activities and salient activities automatically and unsupervisedly. An unified energy function based on the discovery framework was designed to perform anomaly detection at different levels. Experiments on the published UCSD anomaly detection datasets have showed that the proposed method can detect and locate various anomalies effectively and the detection results are better than four comparison algorithms: social force, MDT, MPPCA and social force with MPPCA.

6. REFERENCES

[1] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[2] F. Jiang, J. Yuan, S. Tsaftaris, and A. Katsaggelos, "Video anomaly detection in spatiotemporal context," in *Proceedings of the International Conference on Image Processing*, 2009, pp. 705–709.

[3] B. Morris and M. Trivedi, "Learning, modeling, and classification of vehicle track patterns from live video," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 3, pp. 425–437, 2008.

[4] E. Andrade, S. Blunsden, and R. Fisher, "Modelling crowd scenes for event detection," in *18th International Conference on Pattern Recognition (ICPR)*, vol. 1. IEEE, 2006, pp. 175–178.

[5] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 935–942.

[6] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 2921–2928.

[7] C. Liu *et al.*, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[8] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[9] Y. Rubner, C. Tomasi, and L. Guibas, “The earth mover’s distance as a metric for image retrieval,” *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[10] Q. Zhao, Z. Yang, and H. Tao, “Differential earth mover’s distance with its applications to visual tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 274–287, 2010.

[11] S. Salvador and P. Chan, “Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms,” in *16th IEEE International Conference on Tools with Artificial Intelligence*. IEEE, 2004, pp. 576–584.

[12] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 1975–1981.