An Object-aware Anomaly Detection and Localization in Surveillance Videos

Xianghao Zang, Ge Li, Zhihao Li, Nannan Li, Wenmin Wang
Digital Media R & D Center, Peking University Shenzhen Graduate School
xhzang@sz.pku.edu.cn, gli@pkusz.edu.cn, zhihaoli@sz.pku.edu.cn, linn@pkusz.edu.cn, wangwm@ece.pku.edu.cn

***Abstract*—Abnormal event detection plays an important role in video surveillance and smart camera systems. Existing methods in the literature are usually not object-aware, where different objects are not distinguished in processing. In this work, we propose an efficient object-aware anomaly detection scheme, specifically focusing on certain object categories, such as pedestrians. We first perform a block-based foreground segmentation to confine our analysis to moving objects and avoid irrelevant background dynamics. Then we discard uninterested objects by running an object detector on connected blocks. Finally we extract histograms of block-motion trajectories and cluster them to represent normal events. Our experiments demonstrate the accuracy and efficiency of the proposed method on dataset (PKU-SVD-B). We also propose a clip-based evaluation criterion with practical consideration and discuss this method at last.**

***Keywords*-anomaly detection; surveillance videos; object-aware;**

I. Introduction

Abnormal event detection is one of critical tasks based on videos. This catalyzes important research in computer vision, aiming to find abnormal events automatically [1], [2], [3], [4], [5]. Depending on the particular applications, especially for the police, the interest targets are often on a specific class of objects, such as pedestrians or vehicles. Unfortunately, anomaly detection based on interest objects has not attracted enough attention.

Many anomaly detection techniques have been proposed. Specifically, method using trajectory extraction [6] is ubiquitous. And method by extracting normal distributions of low-level video feature, such as mixtures of dynamic textures (MDT) [7] are also proposed. Though these show comparative improvements over earlier techniques, these low-level methods will put a large computation burden for large-resolution videos.

To suppress undesirable background redundancy and adapt to increasing camera resolution, we present a block-based object-of-interest detection. For example, using a block size of $60*60$, a picture of resolution $1920*1080$ can be divided to 576 blocks. This is illustrated in Fig. 1.

Based on the above method, this paper proposes a unified framework with four steps, shown in Fig. 2. First, foreground objects are considered to be dynamic, and stationary objects to be background. Thus, we take an object as foreground only when it starts to move. Second, not all of the foreground objects are interested for detection in some surveillance applications. As a result, we discard uninterested objects, and extract interest objects for further detection. Third, a trajectory-based method is performed to track these foreground blocks of the interest object. At last, K-means algorithm based on partition-trajectory histogram is used for final abnormal event detection. For more practical evaluation, we also propose a new criterion for anomaly detection.

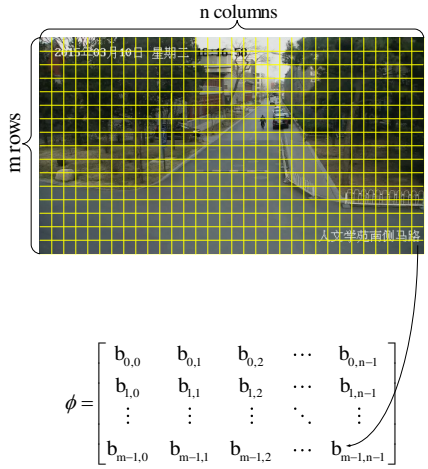


Figure 1. **Block-based representation.** We divide one frame into blocks and each block is represented by summation of pixel values. The matrix ϕ with smaller dimensions is obtained for anomaly detection.

The remainder of this paper is organized as follows. In section 2, the related works are reviewed. In section 3, deep foreground segmentation method including block-based foreground segmentation and patch redundancy elimination is presented. In section 4, feature vector construction for interest objects and anomaly detection are described. Experimental results are discussed in section 5, and we conclude this paper in section 6.

II. Related work

Research in video surveillance has made great progress in recent years. Especially, a lot of works about video anomaly detection and localization have been published, such as object track, action recognition and anomaly inference. To detect anomalies from scenes, Mahadevan *et al.* [2] propose a mixtures of dynamic textures (MDT) model [7] to detect temporal and spatial abnormalities. Kratz *et al.* [5] extract

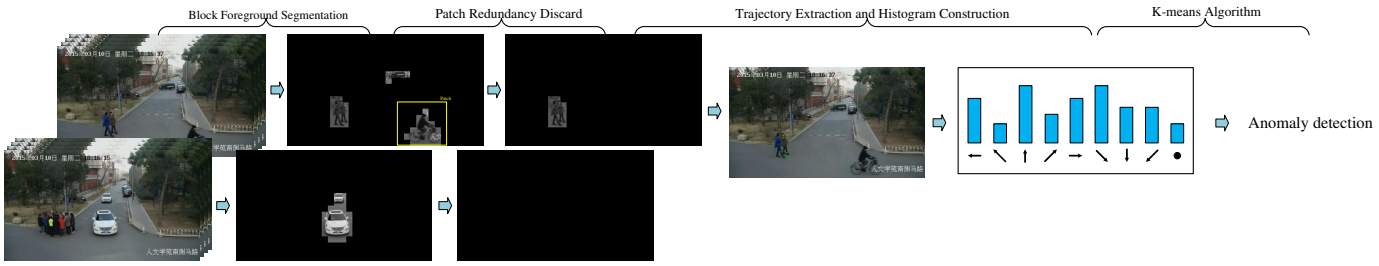


Figure 2. **Block-based framework for anomaly detection**, including foreground segmentation, elimination of redundant patch, trajectory extraction, trajectory-histogram generation, and anomaly detection. It suppresses undesirable background redundancy and adapts to increasing camera resolution with relatively low complexity.

spatio-temporal gradients to fit Gaussian models, then use HMM to detect abnormal events. Since they analyze difference between current and past features, these approaches are time-consuming.

As for modeling group interaction events, Mehran *et al.* [3] present a new way to formulate the abnormal crowd behavior by adopting the social force model, and then use Latent Dirichlet Allocation (LDA) to detect anomalies. Cui *et al.* [4] propose a method based on potential energy interaction to model the inter-personal relationship. These models strongly adhere to motion information and are limited in specific scenarios.

III. Deep Foreground Segmentation

In this section, we describe our deep foreground segmentation algorithm in detail.

A. Block-based Foreground Segmentation

First, we represent the input frame using a block-based method. We get a grey frame from the input image. Set the resolution of one frame : $P * Q$, the size of each block : $p * q$. The grey one has $m * n$ blocks where $m = P/p$ and $n = Q/q$. For each block, we define the feature value $b_{s,t}$ as Eq. 1.

$$b_{s,t} = \sum_{i=1}^{p} \sum_{j=1}^{q} p_{i,j}, \tag{1}$$

where $b_{s,t}$ and $p_{i,j}$ are feature value and pixel value of the block of s row and t column in grey frame respectively. Now, we get feature matrix ϕ in Fig. 1.

Second, we obtain the foreground block based on the difference among adjacent frames. For i^{th} and $(i-1)^{th}$ frames, we set feature matrix ϕ_i and ϕ_{i-1}. Threshold ε_1 is predefined to measure the change between the corresponding blocks $b_{s,t}^{\phi_i}$ and $b_{s,t}^{\phi_{i-1}}$ from ϕ_i and ϕ_{i-1} frames respectively.

Considering the effect of dynamic background, set ψ as background matrix. It is updated using the average of blocks in last video clip Ψ, in which the number of frames ϕ is ι.

$$b_{s,t}^{\psi} = \frac{1}{\iota} \sum_{i=1}^{\iota} b_{s,t}^{\phi_i} \quad (\phi_i \in \Psi). \tag{2}$$

Where

$$|b_{s,t}^{\phi_i} - b_{s,t}^{\phi_{i-1}}| < \varepsilon_1 \quad (i \in [1, \iota]) \tag{3}$$

Another threshold ε_2 is predefined to detect the foreground block. A new matrix $\boldsymbol{\varphi}$ is constructed to record the $b_{s,t}^{\phi_i}$ using the Eq. 4. Therefore, we obtain sparse matrix $\boldsymbol{\varphi}$:

$$\varphi_{s,t} = \begin{cases} b_{s,t}^{\phi_i} & |b_{s,t}^{\phi_i} - b_{s,t}^{\phi_{i-1}}| \geq \varepsilon_1 \ \cup \ |b_{s,t}^{\phi_i} - b_{s,t}^{\psi}| \geq \varepsilon_2 \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

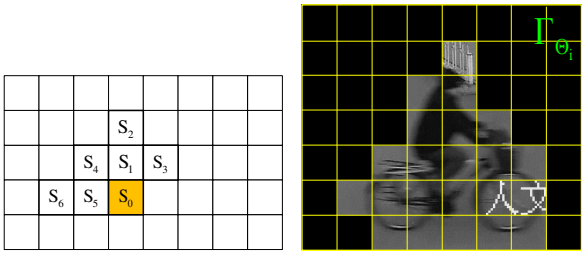
Third, by reducing the average of pixel values from original grey frame, we eliminate the side effects from the block-based method due to instantaneous illumination variation.

As a result, we get a sparse matrix $\boldsymbol{\varphi}$, which represents the foreground for further process.

B. Elimination of Redundant Patch

In this section, we further eliminate the redundant blocks in the matrix $\boldsymbol{\varphi}$ by using object-aware method. Without loss of generality, pedestrians are set as the only interest object.

First, we pack the foreground blocks $\varphi_{s,t}$ into several groups. As illustrated in Fig. 3(a). We traversal all the blocks in sparse matrix $\boldsymbol{\varphi}$ in row by row. In order to create a group Θ, we define position $S_0 = \varphi_{s,t}$.



(a) order of blocks to be searched (b) rectangle picture

Figure 3. (a) the corresponding positions with $S_0, S_1, \ldots, S_6$. (b) Expanding rectangle picture with one block can improve the recognition rate.

When we arrive foreground block S_0 in $\boldsymbol{\varphi}$, we search $S_1, \ldots, S_6$. If at least one of them is foreground block, then

S_0 will be allocated into existing group Θ that S_i belongs to. Otherwise, we create a new group.

Second, We generate rectangles according to those groups obtained above. There are many $\varphi_{s,t}$ in each group, and we use them to define the rectangle Γ_{Θ_i} for group Θ_i as illustrated in Fig. 3(b).

At last, we apply DPM [8] on Γ_{Θ_i} to detect pedestrians. We get the fine foreground matrix $\tilde{\boldsymbol{\varphi}}$ by updating each element $\tilde{\varphi}_{s,t}$ as following:

$$\tilde{\varphi}_{s,t} = \begin{cases} \varphi_{s,t} & \varphi_{s,t} \in \text{pedestrian} \\ 0 & \text{Otherwise} \end{cases} \tag{5}$$

Now We can focus our attention on detecting the abnormal events based on $\tilde{\boldsymbol{\varphi}}$, which is only pedestrian-aware.

IV. Abnormal Event Detection

In this section, we first propose a method based on partition-trajectory histogram to represent the spatio-temporal feature of human behaviours. Then we apply K-means algorithm to detect the abnormal events.

A. Partition-Trajectory Histogram

First, we track every foreground block $\tilde{\varphi}_{s,t}$ to get a trajectory. The approach based on dense trajectory can achieve good results for action recognition [6]. We utilize the similar method for each block $\tilde{\varphi}_{s,t}$. Assume the corresponding coordinate of block $\tilde{\varphi}_{s,t}$ is $(\omega_0, \upsilon_0, \omega_1, \upsilon_1)$. With every block, we search for the best match one in next frame using Novel Three Step Search (NTSS) [9]. After this step, we record the position $(\omega_0^1, \upsilon_0^1)$ of the match one and use the match one for next NTSS. This process continues for λ frames.

Second, we divide these trajectories into different partitions, and construct histogram for modeling normal events. We accumulate the extracted trajectories into a histogram with nine bins according to their respective directions. The direction ζ_d of the trajectory is determined by

$$\zeta_d = (\upsilon_0^\lambda - \upsilon_0)/(\omega_0^\lambda - \omega_0). \tag{6}$$

The weight ζ_w can be calculated by

$$\zeta_w = \frac{1}{\lambda}\sqrt{(\upsilon_0^\lambda - \upsilon_0)^2 + (\omega_0^\lambda - \omega_0)^2}. \tag{7}$$

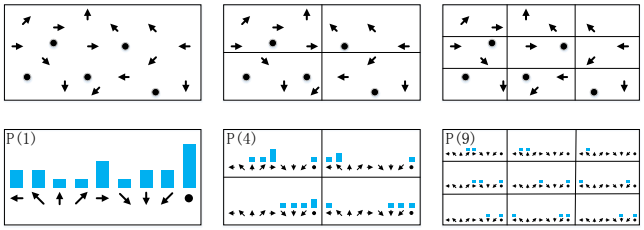


Figure 4. **Partition-Trajectory Histogram.** Each arrow shows the direction of one trajectory. For each partition, we obtain a histogram. Therefore one frame can be represented by a vector of various dimensions.

For the i^{th} bin , we calculate its weight h_i as following:

$$h_i = \sum \zeta_w \quad (\zeta_w \in \text{bin } i). \tag{8}$$

Each frame can be represented by one vector, or the concatenation of four or nine vectors, depending on the number of partitions in a frame, which is illustrated in the second row of Fig. 4.

B. Abnormal Events Detection

In this section, we apply K-means based algorithm to obtain the detection result.

First, We divide one frame into four partitions and construct feature vector with 36 dimensions for representation. After K-means clustering, we use the parameter th_0 to measure the distance between the centroid and feature vector.

Second, we get the abnormal clip $[\rho_0, \rho_1]$ for each video. After the process above, detection result may have several clips which have continuous abnormal frames. We choose the longest one and record its start frame ρ_0 and end frame ρ_1 as the detection result.

V. Experiments

The dataset we take experiments on is the PKU-SVD-B, which includes 360 video clips for train and 120 clips for test. And each video clip lasts for 30 seconds. Frame rate is 25fps or 30fps. The resolution of this dataset is 1920*1080, which is 10 times larger than dataset Subway[1] that is the largest one of the popular anomaly datasets. The abnormal events only have relationship with pedestrians. There are five kinds of events: cluster, flee, chase, abandon, linger in this dataset.

A. Evaluation

The principal quantitative measure used is the F-score. For each video, the detection result is $[\rho_0, \rho_1]$, which denotes the start frame and end frame. While the ground truth is $[\varrho_0, \varrho_1]$. To be considered as a true positive detection, the length of overlap must exceed 50%

$$TP = \begin{cases} 1 & \frac{[\rho_0,\rho_1]\bigcap[\varrho_0,\varrho_1]}{[\rho_0,\rho_1]\bigcup[\varrho_0,\varrho_1]} \geq 50\% \\ 0 & \text{Otherwise} \end{cases} \tag{9}$$

As we know, abnormal event is usually composed of continuous frames, while detection result comprising many noncontinuous frames has little practical significance. This is the reason why we propose this new criterion. Different from the previous evaluation method, the new detection criterion is clip-level instead of frame-level or pixel-level.

B. Analysis

According to the Eq. 9, we make a obvious conclusion below:

$$TP = \begin{cases} 0 & l_\rho < l_\varrho/2 \ \cup \ l_\rho > 2l_\varrho \\ 1 & \text{Otherwise} \end{cases} \tag{10}$$

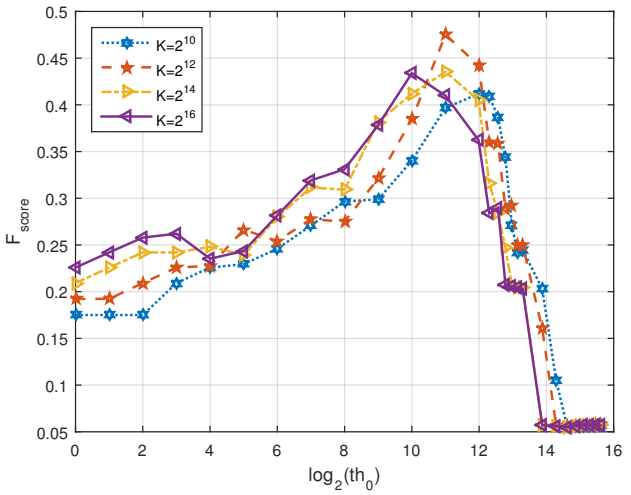


Figure 5. $\mathbf{F_{score}}$ **changes over** $\mathbf{th_0}$. The parameter K represents the different number of clustering centroids. x axis presents the logarithm of th_0.

In which $l_\rho = \rho_1 - \rho_0$, $l_\varrho = \varrho_1 - \varrho_0$. Therefore, the shape of F-score curve is very different from previous work of others.

Through amounts of experiments, we know that curve variation trend becomes stable when $th_0 \geq 2048$. Therefore, we get the best result $F_{score} = 0.475$ when $K = 2^{12}$ and $th_0 = 2048$ as illustrated in Fig. 5. We show some experimental results in Fig. 6.

(a) abandon (b) cluster

(c) linger (d) chase

Figure 6. **Some results from PKU-SVD-B**, Our method is able to detect anomalies, which have anomalous speed and few moving within a period of time.

VI. Conclusion

In this paper, we propose a uniform framework for object-aware anomaly detection. Inside, we introduce a new deep foreground segmentation and a new partition-trajectory histogram algorithms. Since the proposed framework is object-aware, more unrelated objects are eliminated before anomaly detection. As a result, this framework gives a relatively low computation complexity and high detection accuracy. Meanwhile, for practical consideration, we also propose a reasonable evaluation criterion. We analyze it experimentally and obtain a best F-score through amounts of comparison experiments.

Acknowledgment

This work was partly supported by the grant from Shenzhen municipal government for basic research on Information Technologies (No. JCYJ20130331144751105), and Shenzhen Peacock Plan. The author also thanks National Engineering Laboratory For Video Technology (NELVT), Peking University for providing anomaly detection dataset (PKU-SVD-B).

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 3, pp. 555–560, 2008.

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

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

[4] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3161–3167.

[5] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1446–1453.

[6] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[7] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 5, pp. 909–926, 2008.

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[9] J. Y. Tham, S. Ranganath, M. Ranganath, and A. A. Kassim, "A novel unrestricted center-biased diamond search algorithm for block motion estimation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, no. 4, pp. 369–377, 1998.