MULTI-SCALE ANALYSIS OF CONTEXTUAL INFORMATION WITHIN SPATIO-TEMPORAL VIDEO VOLUMES FOR ANOMALY DETECTION

Nannan Li[1], *Huiwen Guo*[1], *Dan Xu*[2], *Xinyu Wu*[1,2]

[1]Guangdong Provincial Key Lab of Robotics and Intelligent System,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2]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 scenes. The proposed approach detects anomalies based on the contextual information analysis within spatio-temporal video volume. Around each pixel, spatio-temporal volumes are built and clustered to construct the activity pattern codebook. Then, the composition information of the volumes within a large spatio-temporal window is described via a dictionary learned by sparse representation. Furthermore, multi-scale analysis is employed to adapt the size change of abnormal events. Finally, the sparse reconstruction cost is designed to evaluate the abnormal level of an input motion pattern. We demonstrate the efficiency of the proposed method on the existing public available anomaly-detection datasets and the performance comparasion with three existing methods validates that the proposed method detects anomalies more accurately.

***Index Terms*—** video anomaly detection, contextual information, bag-of-features, sparse representation

1. INTRODUCTION

Video anomaly detection has been a popular research area of intelligent visual surveillance due to the increasing security needs. Although the practical application of the research has a captivating prospect, there are two difficulties making it an challenging problem. One is the ambiguity of anomaly definition, since we do not know what kinds of anomalies will appear in a given scene; the other is the complexity of scenes due to the change of crowd density, varying perspectives and occlusion. There is not an all-purpose method suitable for every circumstance.

Generally, the approaches focused on this aspect can be divided into two categories. One of them is based on trajectories [1][2][3]. The main idea of such methods is to construct activity model from trajectories collected by tracing individual moving object in the video, then to detect objects with abnormal trajectories as anomalous ones. Those methods perform efficiently, if we can obtain the clear trajectories of objects. Many researchers propose various methods to address the tracking problem, such as Song et al. apply robotic PTZ cameras assisted by a wide-angle camera to obtain trajectories of multiple person for autonomous and scalable crowd surveillance [4]. However, in dense crowds, to achieve robust tracking is quite difficult because of serious occlusion problems, which heavily degrades the performance of the anomaly detection [5]. The other category of approaches address the problem by describing activity patterns based on low-level features. It extracts features representing appearances or motion from the video, then builds a probabilistic model to describe normal activity patterns, and activities with a low probability value under the model are considered as anomalies. Kratz et al. propose to model the crowd scenes via a coupled Hidden Markov Model using spatio-temporal gradient as feature representation [6]. Mahadevan et al. use a mixture dynamic texture model combined with a salient patch detecting method for spatial and temporal anomalies detection [7]. Considering the interaction between local activity, Kim et al. apply a spatio-temporal Markov Random Field model to detect local abnormal activity through a global inference [8]. Inspired by the social behavior research, Mehran et al. employ Latent Dirichilet Allocation model [9] to represent the force flow distribution of normal activity pattern for global unusual event detection [10]. However, these methods model activity patterns only considering global or local context on single image frame, neglecting the contextual information of spatio-temporal video volumes(STVVs), which may be critical for anomaly detection in some cases, i.e. detection abnormal events occurring not in a normal order.

In this paper, we aim to detect anomalies considering the contextual information of STVVs. To achieve this task, STVVs around each pixel are constructed and represented with learned activity pattern codebook through clustering. The composition information of STVVs in a large spatio-spatio video window is described through the improved bag-of-features method. Then the sparse representation method is used to build the dictionary of composition patterns. Finally, we design a sparse reconstruction cost to quantify the normalness of events. Our work is inspired by [11], where the author apply a probabilistic framework to describe the composition relationship among STVVs. The overview of our proposed approach is illustrated in Fig. 1.

2. ACTIVITY PATTERN CODEBOOK CONSTRUCTION FOR SPATIO-TEMPORAL VOLUME

2.1. Feature representation of spatio-temporal volume

To extract low-level features for activity pattern codebook construction, we densely sample the video, constructing a STVV of size $L_x \times L_y \times L_t$ around each pixel (where $L_x \times L_y$ is the spatial size of

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).

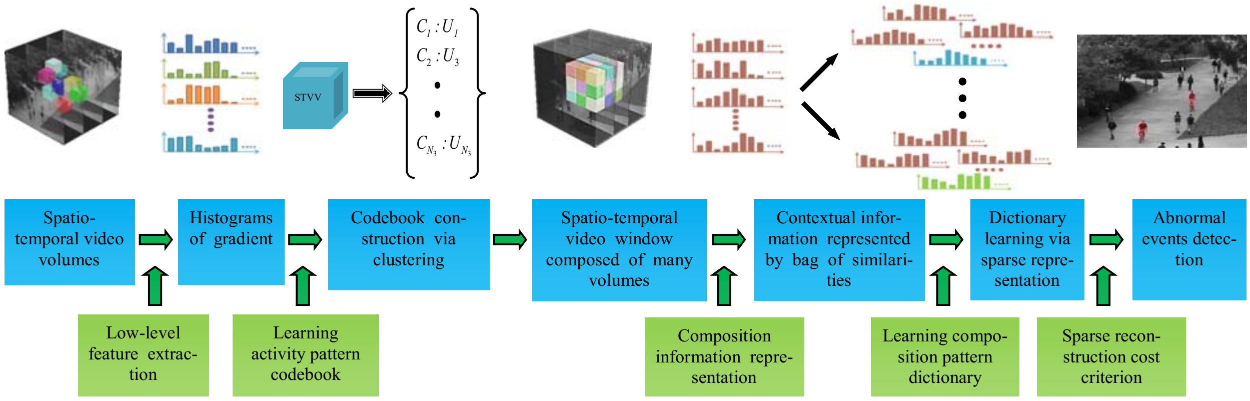
STVV
$C_1 : U_1$
$C_2 : U_3$
$C_{N_3} : U_{N_3}$
Spatio-temporal video volumes
Histograms of gradient
Codebook construction via clustering
Spatio-temporal video window composed of many volumes
Contextual information represented by bag of similarities
Dictionary learning via sparse representation
Abnormal events detection
Low-level feature extraction
Learning activity pattern codebook
Composition information representation
Learning composition pattern dictionary
Sparse reconstruction cost criterion

Fig. 1: The overview of the proposed approach for anomaly detection

STVV and L_t is the depth of STVV in time). Within each STVV, we calculate spatio-temporal gradient at each pixel. Assigning that G_x and G_y are the spatio gradients and G_t is the temporal gradient, then the STVV can be represented by the histogram of spatio-temporal gradients in polar coordinates as follows:

$$M_s = \sqrt{(G_x^2 + G_y^2)},$$

$$[M_{st}, \theta, \phi] = [\sqrt{(M_s^2 + G_t^2)}, \arctan(G_y/G_x), \arctan(M_s/G_t)], \quad (1)$$

where M_s represents the magnitude of spatial gradient. We compute G_x, G_y, G_t using one order difference approximation. Before calculating the gradients, we filter the image via a Gaussian kernel with bandwidth σ(typically 0.5) to suppress the noise. The range of θ and ϕ are set as [-π,π] and [-π/2,π/2] and are quantized in eight and four bins, respectively. The description of STVV is constructed as a 12-dimension vector, which is obtained by accumulating magnitude of gradients in the 12 bins. Then we normalize the feature vector to eliminate the magnitude order difference among vector dimensions. The normalized feature vector is denoted as h_j.

2.2. Activity pattern codebook construction via clustering of STVVs

We obtain the feature vector, h_j, for the spatio-temporal volume, v_i. These feature vectors represent motion patterns at STVVs, which are used to construct activity pattern codebook. We build a codebook that represents the distribution of the descriptor vectors. Since the number of vectors is very large, we choose the weighted fuzzy c means (WFCM) algorithm proposed by Hore et al. [12] for codebook construction because of its ability to process data by chunk. The main idea is to consider a chunk of data, cluster it, then the data is condensed into several center points, therefore addressing the limited memory problem. Finally, these center points are clustered to obtain the codebook. The objective function of weighted fuzzy c means clustering for our task is written as: $J = \Sigma_{i=1}^{N_c} \Sigma_{j=1}^{N_h} u_{i,j}^m \omega_j D_{i,j}(h_j, c_i)$, where $u_{i,j}^m$ is the similarity matrix, ω_j is the weight of the sample, $D_{i,j}(h_j, c_i)$ is the similarity measurement between h_j and center c_i, using Euclidean distance. The details of codebook construction via WFCM clustering are illustrated in Algorithm 1.

Hore et al.[12] prove that if we set the number of clusters bigger or equal to the real number of clusters, there is no information loss at each chunk clustering. Here we let the number of clusters, N_c, equal to 40, the maximum number of clusters that may exist in our case. Then, each STVV, h_j, will be represented by a set of similarity values, $\{u_{i,j}\}_{i=1}^{N_c}$.

3. DESCRIPTION OF CONTEXTUAL INFORMATION WITHIN SPATIO-TEMPORAL WINDOW

Bag-of-features is a popular method used for describing the contextual information of an image. The main drawback of such method is that it neglects the spatial composition relationship of features. Lazebnik et al.[13] propose an improved method. Here, we extend their method from two dimensional to three dimensional to represent the composition relationship of STVVs. This operation can be seen as a tradeoff between subdividing and disordering. we select a large spatio-temporal video window that contains many STVVs around each pixel. We divide the spatio-temporal window into eight blocks evenly. Within each block, we calculate the bag of similarities as follows: using WFCM algorithm, we evaluate the similarities of each STVV to the N_c clusters, then accumulate all the similarities corresponding to the same cluster respectively to form a histogram vector. Considering similarities as features, the process is the standard trick of bag-of-features. Having obtained the histogram vectors in each block, we construct the composition representation vector as follows: the bins of the same position from each histogram are arranged together to form the vector fragment, then all the vector fragments are concatenated to construct the composition representation vector. Although the operation seems trivial, it possesses merits from two aspects. On one hand, the bag of similarities method, which takes the form of the accumulation of local similarities, has the performance for expressing local context in each block. On the other hand, the representation vector constructed by concatenating bag of similarities vectors, captures the composition information within a spatio-temporal window. A simple example of constructing vectors representing composition information in two-dimensional case is illustrated in Fig. 2.

Algorithm 1 Cluster the feature vectors via WFCM

```
Input:
    Feature vectors $h_j$ from the entire video.
Output:
    Cluster centers $c_i$, $i = \{1, 2, ...N_c\}$.
Divide $h_j$ into $k$ chunks, assign that $kTh$ chunk has $N_k$ vectors.
for $t = 1$ to $k$ do
    Load in the memory the $tTh$ chunk of data: $h_j^t$, $j = \{1, 2, ..., N_t\}$; initialize $c_i^t$ via applying k-means cluster upon $h_j^t$, and let $\omega_j = 1$; then solve the problem via the algorithm proposed by [12]:
    $\min_{c_i^t} \sum_{i=1}^{N_c} \sum_{j=1}^{N_t} u_{i,j}^m \omega_j D_{i,j}(h_j^t, c_i^t)$
        where $u_{n,j} = (\sum_{i=1}^{N_c} (\frac{\|h_j^t - c_n^t\|}{\|h_j^t - c_i^t\|})^{\frac{2}{m-1}})^{-1}$, $c_i^t = \frac{\sum_{j=1}^{N_t} \omega_j u_{i,j}^m h_j^t}{\sum_{j=1}^{N_t} \omega_j u_{i,j}^m}$,
        $D_{i,j}(v_j^t, c_i^t) = \|h_j^t - c_i^t\|_2^2$
    Let $\omega_i^t = \sum_{j=1}^{N_t} u_{ij}$; Save $c_i^t$ and $\omega_i^t$
end for
Load in memory $c_i^t$ and $\omega_i^t$, $t = 1, 2, ..., k$, $i = 1, 2, ..., N_c$;Denote the ensemble of $c_i^t$, $\omega_i^t$ as $\bar{c_j}$ and $\bar{\omega_j}$ respectively, $j = \{1, 2, ...k \times N_c\}$, then solve the problem: $\min_{c_i} \sum_{i=1}^{N_c} \sum_{j=1}^{k \times N_c} u_{i,j}^m \bar{\omega_j} D_{i,j}(\bar{c_j}, c_i)$
return a set of cluster centers $c_i$, $i = \{1, 2, ..., N_c\}$
```

4. DICTIONARY LEARNING OF COMPOSITION PATTERN AND ANOMALY DETECTION

Using the representation vectors obtained from section 3, we construct a model to represent the composition pattern for each pixel. As the dimension of representation vector, *d*, is a large number (e.g. 320), with a small number of training samples, it is difficult to conduct density estimation. Since sparse representation is suitable to represent high-dimension vector, we propose to construct a sparse dictionary to represent the composition pattern. Through building a normal basis set, the normal events can be reconstructed via the linear combination of the basis set sparsely with a less reconstruction cost, while abnormal events may generate a dense representation with a large reconstruction cost. The objective function of dictionary learning can be written as follows:

$$\min_{D \in R^{d \times k}, \alpha \in R^{k \times n}} \frac{1}{n} \sum_{i=1}^{n} (\frac{1}{2}\|x_i - D\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1), \qquad (2)$$

where x_i is the representation vector, $i = 1, 2, ...n$; D is the objective dictionary to be learned, the size of which is k , and it satisfies that: $\|D_{.j}\|_2 = 1, s.t. \forall j = 1, ..., k$; α_i is the reconstruction coefficient corresponding to x_i; $\|\alpha_i\|_1$ is L_1 norm. The approach to solve problem (2) is to alternate between the two variables, D and α_i, minimizing over one while keeping the other one fixed, as proposed by Mairal et al.[14]. Although it is an online dictionary learning method, it leads to dictionaries with better performance than that resulted from classical batch algorithms for our case. After constructing the dictionary D, given a vector x, the reconstruction coefficient α^* is calculated as follows:

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1, \qquad (3)$$

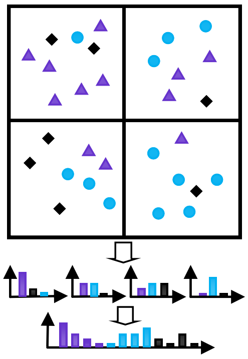

Fig. 2: The simle example of representation vector construction in two-dimensional case. A large region is divided into four small regions, within which standard bag-of-features vectors are constructed. Then those vectors are concatenated to form the representation vector.

then we evaluate the sparsity reconstruction cost (SRC) of x by: $C_x = \frac{1}{2}\|x - D\alpha^*\|_2^2 + \lambda\|\alpha^*\|_1$. The value of SRC, C_x reflects the abnormal level. A high SRC value signifies that the composition pattern represented by the tested vector has a high probability to be abnormal. Then for a given SRC threshold C_{th}, if $C_x > C_{th}$,this vector is classified as an abnormal one.

5. EXPERIMENT VALIDATION

5.1. Multi-scale analysis of video

The proposed algorithm is evaluated in the public available anomaly detection dataset from University of California San Diego (UCSD) [7]. It consists of two subsets, ped1 and ped2, with the resolution of 238 × 158, 360 × 240, respectively. The density of people varies from very sparse (e.g. several people)to dense (e.g. dozens of people), and normal behaviors present in the video are people walking alone or in groups. The abnormal events include: bicyclers, vehicles, skateboarders, people waking on the lawn etc. As mentioned previously, we densely sample the video. The size of STVVs around each pixel is set as 5. Since the size of objects change due to their distances from the camera, besides, we are not acquainted with objects of what size will generate anomalies, it is necessary to analyze the video at multiple scales. But the high computational overhead of a dense sampling in scales is too expensive, in our experiments the scale of spatio-temporal window is restricted to two different levels, 17 × 17 × 17 and 33 × 33 × 33. Experiments demonstrate that there is a higher detection accuracy but more false positives via a small spatio-temporal window, while for a large window, the result is the opposite. The final detection result is obtained from the intersection of that two results. This operation filters out the spurious false positive detections and increases the accuracy of localizing abnormal objects. From a probabilistic perspective of view, each detection result of the two levels independently represents how likely an observed activity pattern, within a spatio-temporal window, is an outlier; the final result is produced via a product rule, resulting in the spatial intersection of the two detected results, following the processing proposed by [15]. Examples of anomaly localization results at two different levels are illustrated in Fig. 3.

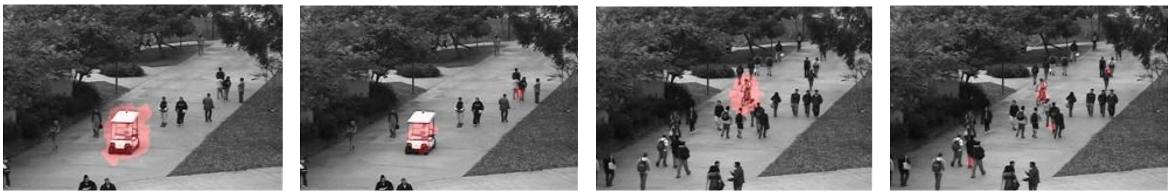

Fig. 3: Examples of anomaly localization in Ped1 Dataset. The columns of 1 and 3 are the results of coarse scale, while the columns 2 and 4 are those of fine scale.

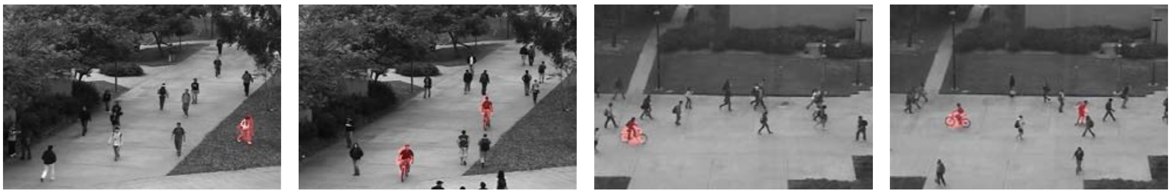

Fig. 4: Examples of anomaly detection results with the proposed approach on Ped1 Dataset(the front two columns) and Ped2 Dataset(the back two columns).

5.2. Performance evaluation

For each video clip, the frame-level groundtruth uses a binary flag to indicate whether one or more anomalies are present in the current frame. To evaluate the performance of the proposed algorithm, we conduct experiments on ped1 and ped2 with different thresholds based on the frame-level groundtruth. To compare performance, we choose three state-of-the-art methods including the Mixture of Dynamic Textures (MDT) [7], the Social Force [10] and the optical flow monitoring method proposed by Adam et al. [16]. Fig. 5 illustrates the ROC curves of different methods on Ped1 and Ped2, respectively. We calculate the area under ROC curve (AUC), and the average AUC of our algorithm on the two datasets is 86.2%, which is better than that of MDT, 82.4%, the best one of three approaches for comparison. The proposed algorithm is implemented with Matlab. Under a standard PC platform with Intel core2 Duo and 3 GHz CPU and 2 GB memory, the processing time of our algorithm is 5.7 seconds per frame. Since the similarity matrix computation is fast and the composition information of spatio-temporal window can be obtained via only altering accumulation from two layers of STVVs, the main time overhead is from the evaluation of sparse reconstruction cost. Examples of anomaly detection results are illustrated in Fig. 4.

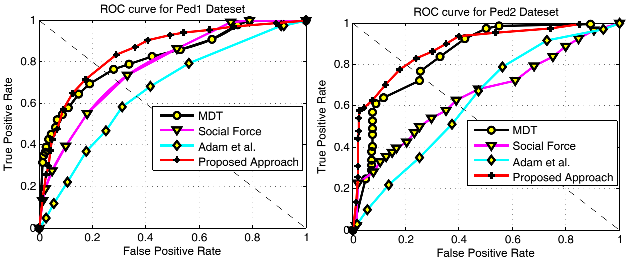

Fig. 5: ROC curves of different approaches on Ped1 and Ped2 Dataset.

6. CONCLUSION AND FUTURE WORK

This paper presented a novel approach for anomaly detection in crowded scenes. This approach was based on the analysis of the contextual information within spatio-temporal video volume. Around each pixel, spatio-temporal video volumes were constructed, then clustered for building activity pattern codebook. A dictionary representing composition information of the volumes within a large spatio-temporal window was learned via sparse representation. Finally a spare reconstruction cost criterion was designed to perform anomaly detection. Future research will extend the approach by using online learning method to make the codebook and dictionary update according to new input video streams.

7. REFERENCES

[1] C. Stauffer and W. Grimson, “Learning patterns of activity using realtime tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, 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, pp. 425–437, 2008.

[4] Y. Xu and D. Song, “Systems and algorithms for autonomous and scalable crowd surveillance using robotic ptz cameras assisted by a wide-angle camera,” *Autonomous Robots*, vol. 29, pp. 53–66, 2010.

[5] D.Xu, X. Wu, D. Song, N. Li, and YL. Chen, “Hierarchical activity discovery within spatio-temporal context for video anomaly detection,” in *Proceedings of the International Conference on Image Processing*, 2013, pp. 3597–3601.

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

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

[8] J. Kim and K. Grauman, “Observe locally, infer globally: a spacetime mrf for detecting abnormal activities with incremental updates,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2921–2928.

[9] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

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

[11] Javan Roshtkhari, Mehrsan, and Martin Levine, “An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions,” *Computer vision and image understanding*, vol. 117, pp. 1436–1452, 2013.

[12] P. Hore, L. Hall, D. Goldgof, Y. Gu, A. Maudsley, and A. Darkazanli, “A scalable framework for segmenting magnetic resonance images,” *Journal of Signal Processing Systems*, vol. 54, pp. 183–203, 2009.

[13] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178.

[14] J. Mairal, F. Bach, and J. Ponce, “Online dictionary learning for sparse coding,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[15] M. Bertini, A. Del Bimbo, and L. Seidenari, “Multi-scale and realtime non-parametric approach for anomaly detection and localization,” *Compt. Vis. Image Und.*, vol. 116, pp. 320–329, 2012.

[16] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust real-time unusual event detection using multiple fixedlocation monitors,” *PAMI*, vol. 30, pp. 555–560, 2008.