Abnormal Event Detection using Scene Partitioning by Regional Activity Pattern Analysis

Jongmin Yu, Jeonghwan Gwak, Seongjong Noh and Moongu Jeon
School of Information and Communication, Gwangju Institute of Science and Technology, 61005 Gwangju, Republic of Korea

Keywords: Abnormal Event Detection, Scene Partitioning, Spatio-temporal Feature, Optical Flow.

Abstract: This paper presents a method for detecting abnormal events based on scene partitioning. To develop the practical application for abnormal event detection, the proposed method focuses on handling various activity patterns caused by diverse moving objects and geometric conditions such as camera angles and distances between the camera and objects. We divide a frame into several blocks and group the blocks with similar motion patterns. Then, the proposed method constructs normal-activity models for local regions by using the grouped blocks. These regional models allow to detect unusual activities in complex surveillance scenes by considering specific regional local activity patterns. We construct a new dataset called GIST Youtube dataset, using the Youtube videos to evaluate performance in practical scenes. In the experiments, we used the dataset of the university of minnesota, and our dataset. From the experimental study, we verified that the proposed method is efficient in the complex scenes which contain the various activity patterns.

1 INTRODUCTION

In recent years, to develop practical surveillance systems, researchers have investigated several challenging issues such as the detection of abandoned luggage (Tian et al., 2011) (Bayona et al., 2009) (Pan et al., 2011), the person tracking (Krumm et al., 2000) (Javed and Shah, 2002) (Sato and Aggarwal, 2001) (Liem and Gavrila, 2014) and the unusual event detection (Roshtkhari and Levine, 2013) (Cong et al., 2011) (Mehran et al., 2009) (Zhou et al., 2015) (Li et al., 2014). These studies have provided much helpful functions to the users of surveillance systems. Especially, detection of abnormal events is the one of the important studies for surveillance systems to provide information of unusual or suspicious activities to users. In general, abnormal events are detected through the following procedures. Initially, activity models for normal behaviors are constructed by learning activity patterns of objects observed in normal videos which do not include abnormal events. Next, the abnormality of events starts to be detected by measuring similarity between input data and the learnt models, and distinguish whether there is normal or abnormal events by using a specific threshold.

To implement this strategy in actual surveillance applications, we should consider two major issues. The first is to transform individual or group activities in the scenes into trainable patterns. The surveillance videos contains a lot of activities that are generated by various moving objects. Raw information such as color or gray images is not sufficient to distinguish between normal and abnormal events. Therefore, well-defined features for representation of these activities are required. To transform activities into useful patterns, some researchers used texture information such as a histogram of gradient (HOG) (Dalal and Triggs, 2005), spatio-temporal gradient (Klaser et al., 2008), and mixture of dynamic textures (Mahadevan et al., 2010). Also, spatio-temporal frequency (Wang et al., 2012) has been considered an efficient means of detection. On the other hand, some other researchers have focused on crowd activity modelling based on flow information such as optical flow (Andrade et al., 2006b) and force flow (Mehran et al., 2009).

The second is to model normal events to measure a similarity. The model of normal events is constructed to obtain general patterns for normal event from frame sequences which contains only normal events. Dictionary learning (Cong et al., 2011) and the bag of words (BoW) model (Mehran et al., 2009) are widely used, and probabilistic methods such as hidden markov model (HMM) (Andrade et al., 2006a) are also adopted.

Yu, J., Gwak, J., Noh, S. and Jeon, M.
Abnormal Event Detection using Scene Partitioning by Regional Activity Pattern Analysis.
DOI: 10.5220/0005720606340641
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 3: VISAPP*, pages 636-643
ISBN: 978-989-758-175-5

Figure 1: Right red boxes represent a distortion of flow patterns caused by distance between objects and camera. Left blue boxes show flow patterns of motorcycle and pedestrians.

However, the constructed models of normal events in the previous studies have used whole events that are occurred within a frame without consideration of characteristics of activities. Unfortunately, these approaches are not sufficient to capture the abnormal event in practical situations. Videos obtained from surveillance systems can contain a lot of diverse moving objects, and these moving objects can generate variety activity patterns. For instance, the property of the activity patterns in sidewalks and roads are different. Moreover, according to the geometric conditions, such as angle and shooting distance of a surveillance camera, these activity patterns can be distorted. In other words, these activity patterns can be represented differently according to distance and angle between the camera and objects. Figure 1 shows the diverse activity patterns caused by various moving objects and the geometric condition of the surveillance camera.

To handle these problems, we propose a novel method for detection of abnormal event using scene partitioning. We divide an image into blocks of a uniformed size, and extract spatio-temporal volume from the blocks. Depending on activity patterns of each spatio-temporal volumes, we gather the blocks with similar patterns of activities, and define a region model using the clustered blocks within a frame. By using this region model, the proposed method can consider more practical situations which contain diverse types of activity patterns. The proposed method uses spatio-temporal volumes consisting of dense optical flow (Alvarez et al., 2000). To define the regions, entropy and magnitude of each spatio-temporal volume are used to represent activity patterns of each block. We design a new algorithm based on an expectation-maximization algorithm (EM algorithm) (Moon, 1996) for grouping blocks with similar activity patterns, and we use K-means clustering algorithm to construct the bag of volume (BoV) of each region. The main contribution of this paper is as follow. in order to detect an abnormal event within complex surveillance scenes, we propose anomaly detection method using local-region models which are constructed by scene partitioning.

The rest of the paper is organized as follows. In Section 2, we present the previous works of research on abnormal event detection. In Section 3, we describe the proposed method. In Section 4, the test datasets and experimental results are shown. Finally, the conclusion is presented in Section 4.

2 PREVIOUS WORKS

The detection method for abnormal events, as an important function in video surveillance systems, has provoked a lot of attentions. The research on abnormal event detection has made a lot of progress in recent years. Abnormality of event is detected by means of the likelihood ratio test with normal event model of each abnormal event detection approach. Generally, events that are detected from previous works can be classified into two types of events which is composed of local abnormal events (LAE) and global abnormal events (GAE). LAE is the behaviors of individuals which are different from the neighbors. GAE is defined that the group behavior of the global scene is abnormal (Cong et al., 2011).

To detect abnormality of event from videos, many proposed methods are focused on describing a uniqueness of activities by measures similarity using the general model of normal events. To describe the uniqueness of events within a frame, a lot of method are proposed. Detection methods for abnormality of events in trackable situations are difficult to handle in crowded scenes. Recently, the detection methods using tracking techniques focused on the direct use of motion patterns in an image. (Wang and Miao, 2010) is used Kanade-Lucas-Tomasi (KLT) feature tracker (Lucas et al., 1981) to describe motion patterns of moving objects. (Wang and Miao, 2010) used a historical motion descriptor.

Other researches are focused on the untrackable situations such as an extremely crowded scene. These are focused on the modelling of group behaviors, There are a number of methods that have been developed for GAE detection by modeling crowd activities. (Mehran et al., 2009) represents behaviors of crowd using the social force model (SFM) without any tracking method. SFM estimates the interaction force by computing the difference between the desired veloc-

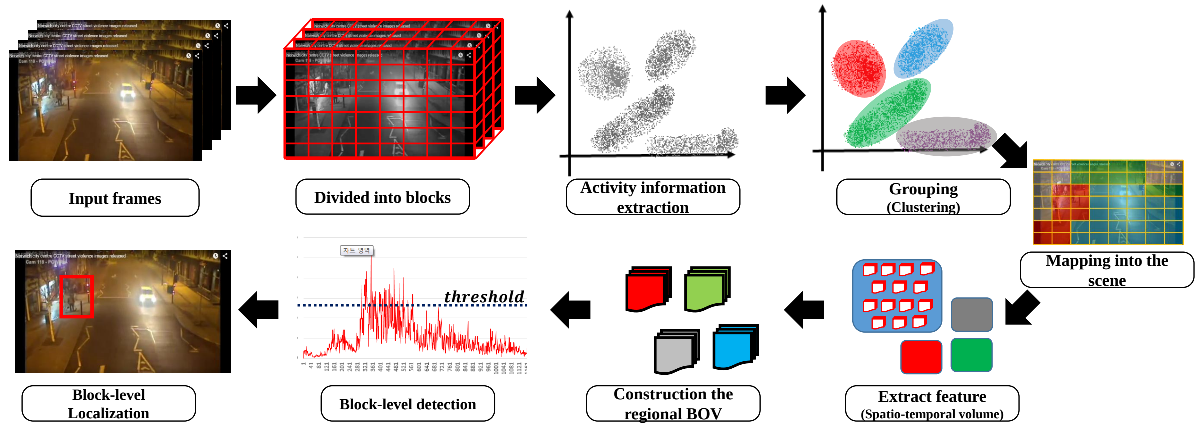


Figure 2: An overall procedure of the proposed method for abnormal event detection and localization in surveillance videos.

ity and the actual velocity. (Cui et al., 2011) models a social behavior using the interaction energy potential.

Meanwhile, the bag of words (BoW) and dictionary learning methods are commonly used to construct the model of normal events. The BoW method was changed to the bag of force (BoF) in (Mehran et al., 2009) to incorporate interaction force. (Andrade et al., 2006a) used the hidden markov model (Eddy, 1996) for analysis of crowd behaviors.

Although these previous methods demonstrated their performance in their own experiments, they mostly focused on either LAE or GAE detection in simple situations. Also, they construct models for normal events using the whole events in the frames. These approaches are inappropriate in modeling various activity patterns which are generated from diverse moving objects and geometrical distortion of a camera. Thus, we argue that modelling of normal events based on the local region model can help to enhance the performance of abnormal event detection.

3 PROPOSED METHOD

In this section, we describe an abnormal event detection and localization method using a scene partitioning based on local activity patterns. In this paper, we focused on the abnormal event detection of complex surveillance scenes through partitioning the frame into several regions and modelling normal event of each region.

Figure 2 illustrates the overall framework of the proposed method. To partition the frame into the set of regions, we divide frames into 3D blocks with a uniform size, and then extract spatio-temporal features from each block sequentially. Based on the block-level activity information, the blocks which have similar activity information are grouped, and form a specific regions within a frame. Based on the region model, the proposed method constructs models of normal activities for each region separately. By using this models, the proposed method can detect region-specific abnormal events.

To distinguish whether events are the abnormal event or not, the proposed method calculates similarity between the regional normal event model and the spatio-temporal feature of blocks of each region. The computed similarities are used to determine anomaly of the events by comparing them with the given fixed threshold. Since the proposed method detect the block-level abnormal events, the method can also localize the block-level position of the abnormal event within the frames.

3.1 Scene Partitioning based on Regional Activity Pattern Analysis

To partition a scene using activity information, we divide the frames into the set of 3D blocks, and we estimate activity patterns of each block indirectly by using the dense optical flow. In order to extract the spatio-temporal features for obtaining training samples, given T frames which only contain the normal events is divided into K clips. Each clip consists of t frames. Each clip is partitioned into uniform-size blocks of $M \times N \times t$. $M \times N$ is the size of the spatial window and t is the depth of the volume in time i.e., frames. In this paper, the temporal depth of the block is equal to the temporal depth of the clip, where these volumes of k-th clip C_k can be described by $\{V_1^k, V_2^k, V_3^k, \ldots, V_n^k\}$, and n is the number of 3D blocks that are extracted from the single clip.

However, raw data of the 3D block can involve redundant information. Therefore, we compute entropy and magnitude from each volume, for removing redundant information and obtaining appropriate information for activities within the volume. The entropy can represent complexity of activity patterns, and the magnitude provides strength of activities such as a motion speed of the corresponding objects. To compute the entropy, we calculate a histogram of optical flow (HOF) (Chaudhry et al., 2009) of each volume. We define the entropy of HOF from each block in k-th clip as follow:

$$S_i^k = -\sum_{j=1}^{B}\left(\frac{h_j^{i,k}}{\sum_{t=1}^{B} h_t^{i,k}} \log \frac{\sum_{t=1}^{B} h_t^{i,k}}{h_j^{i,k}}\right) \quad (1)$$

where S_i^k is the entropy of i-th spatio-temporal volume in the k-th clip, B is the number of bins of the histogram. $h_j^{i,k}$ denotes the number of HOF in the $j-th$ bin of the volume V_i^k. The entropy of HOF is influenced by the diversity of activity patterns. For example, the entropy of scenes which is captured from a highway intersection area is larger than the entropy of a straight highway area. However, the entropy is insufficient to estimate overall properties of activity patterns. In addition, we compute the magnitude of activity patterns within the volumes to represent strength of activities, and it is described as follows:

$$O_{avg}^{i,k} = \frac{1}{H}\sum_{t=1}^{H} O_t^{i,k} \quad (2)$$

where $O_{avg}^{i,k}$ denotes the average magnitude of optical flow of the i-th volume in the k-th clip, H is the number of optical flows in a volume, and $O_t^{i,k}$ is optical flow of the t-th optical flow in the i-th volume within the k-th clip. By using the entropy S_i^k and the magnitude $O_{avg}^{i,k}$, we can represent the activity patterns which is inherent in each volume, and each volume can be defined as

$$V_i^k = \{S_i^k, O_{avg}^{i,k}\} \quad (3)$$

Unfortunately, the entropy and the magnitude, extracted from the single clip, are inappropriate to model the all normal activities that are occured for a long period. Therefore, we normalized the entropies and the magnitudes by computing the expectation of the entropy and the magnitude of all divided clips, and the procedure is describe as follows:

$$\widehat{S_i} = \frac{1}{L}\sum_{l=1}^{L} S_i^l \quad (4)$$

$$\widehat{O^i}_{avg} = \frac{1}{L}\sum_{l=1}^{L} O_{avg}^{i,l} \quad (5)$$

where $\widehat{S_i}$ and $\widehat{O^i}_{avg}$ are the normalized entropy and magnitude of activity patterns for the i-th block. L is the number of clips. After computing the normalized value for entropy and magnitude, the proposed method partitions a scene into several regions which consist of blocks that show similar activity patterns using an algorithm based on the EM algorithm. The algorithm is grouping the blocks using similarity between the entropy and magnitude of each block, the number of centroid is not pre-determined before the training. Fixed number of centroid may cause the over-fitted results in scene partitioning, and thus it can degrade the performance for detecting the abnormal events. We propose an algorithm to increase the number of the centroids int the training phase. Algorithm 1 describes the pseudo code of the proposed algorithm for constructing the region model.

Algorithm 1 : The algorithm for constructing the region model based on the EM algorithm.

Input: $V \longleftarrow$ Set of activity vector of blocks, $\alpha \longleftarrow$ The fixed threshold, $M \longleftarrow$ The number of activity vector of the blocks.

Output: $C_N \longleftarrow$ Centroid of the K-means algorithm, $N \longleftarrow$ The number of centroids of the algorithm $R \longleftarrow$ The label matrix for blocks in frame.

```
C_N is empty set at the beginning of algorithm
C_1 ⟵ V_1
N ⟵ 1
R_1 ⟵ 1
for all i ∈ M do
  for all j ∈ N do
    if argmin_j (Distance(C_j, V_i)) < α then
      Update j-th centroid C_j using V_i
      R_i ⟵ j
    else then
      Create a new centroid and region label
      N ⟵ N + 1
      C_N ⟵ V_i
      R_i ⟵ N
    end if
  end for
end for
```

3.2 Regional Bag of Volumes

The process of identifying the most likely patterns for abnormal events is one of the important issues. In this paper, we construct regional BoVs for representing the general normal activity patterns of each region. Regional BoVs are independently modeled by using the extracted spatio-temporal volume from the 3D block in each region. Through the regional modeling, the regional BoVs can represent more detailed information of motion activities in the regions.

To construct the regional BoV models, we use the K-means clustering. The centroids of the clustering algorithm are used as the codewords. Each centroid is initialized randomly from observed spatio-temporal volumes extracted from each region before training. Then, by measuring the similarity between centroids and observed spatio-temporal volumes, centroids are updated. When training is finished, centroids are used to the codewords of the regional BoV. We use an Euclidean distance for measuring the similarity.

In this process, the number of codewords is important. A large number of codewords can not only increase computation cost, but also can involve redundant information for normal event. On the other hand, too small numbers of codewords can not provide sufficient information to detect abnormal events. For this reason, we determined the scale of the regional BoVs through experimental simulations in this work.

3.3 Abnormal Event Detection and Localization

To detect anomaly from events that are generated from each block of each region, we measure similarity between the input spatio-temporal volume and regional BoV which is modeled from the each region that has the location information of the input block. The proposed method used the Kullback-Leibler (KL) divergence (Goldberger et al., 2003) to compute the similarity. The similarity s_i^t between the spatio-temporal volume of i-th block in region within t-th frame and the regional BoV of the region including that block is describes as follow:

$$s_i^t = \sum_{j=1}^{N} D_{KL}\left(P\left(V_i^t\right)||P\left(M_j\right)\right) \tag{6}$$

where $P\left(V_i^t\right)$ is the probability distribution model which computed from the input spatio-temporal volume of i-th block in the region within t-th frame, and $P\left(M_j\right)$ is a probability distribution model of j-th codeword of regional BoV which is specified by the region. Then, we determine abnormality of activity pattern of each block by comparing it with fixed threshold β. This determination process is described as follow:

if $s_i > \beta$ **then**
  V_i is abnormal
else
  V_i is normal
end if.

By using this the block-level determination process, the proposed method can specify a location which contains the abnormal event. Moreover, by expanding the determination process, the proposed method can detect the frame-level abnormal event.

4 EXPERIMENTAL RESULT

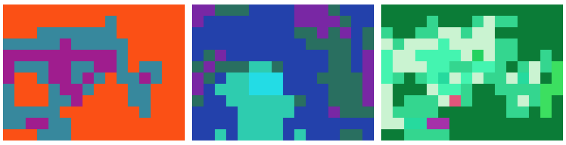
The proposed method was tested on a public dataset and our GIST Youtube dataset in order to demonstrate effectiveness for abnormal event detection. In our first experiments, we used the publicly available dataset of the University of Minnesota. To evaluate the performance, we also compared the results with other abnormal event detection approaches such as pure optical flow, social force model.

In the second experiments, used the GIST Youtube dataset for testing in practical situations. We gathered videos in realistic environment using a semantic image crawler. The dataset is composed of 3 videos which contain real abnormal events.

4.1 UMN Dataset

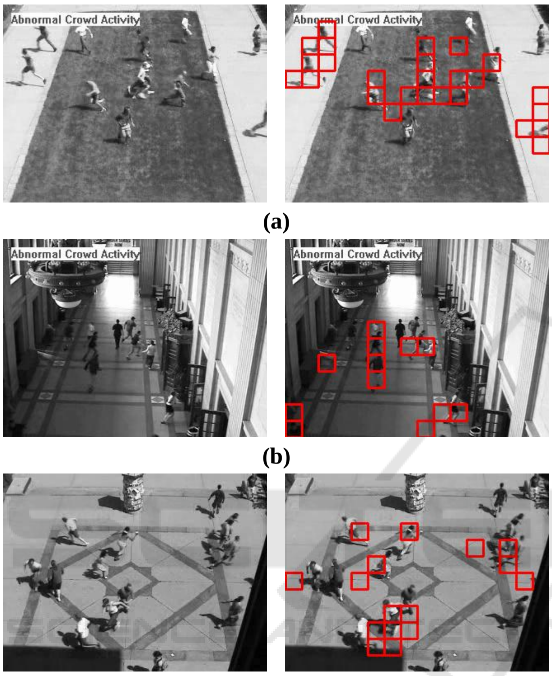
First, we evaluated our method using publicly available dataset which is released by University of Minnesota (called UMN dataset). The UMN dataset consists of 3 different videos of crowd events. Each video contains an escape situation in indoor and outdoor environment. Each video is composed of an initial part of the normal events, and multiple consecutive frames which contain abnormal events. In this experiment, we used initial 200 frames for scene partitioning and constructing regional BoVs.

In the stage of scene partitioning, we first compute the dense optical flow. Next, we divide each frame into 20×20 blocks, and extract $20 \times 20 \times 10$ spatio-temporal volume of optical flow from each block. We found optimal threshold using cross-validation. In conclusion, $\alpha = 0.2$ and $\alpha = 0.105$ are used for video 1 and 2 respectively, and $\alpha = 0.1$ is used for video 3. Figure 3 illustrates results of scene partitioning of each video in UMN dataset. In the experiments, $\beta = 900$ is used, and the size of the regional BoVs is 20. Table 1 provides the quantitative comparison results. The area under ROC (AUC) of our method is 0.9938. The values of AUC demonstrate that the proposed method is better than the methods using the pure optical flow and social force model (SFM) (Mehran et al., 2009). Figure 4 shows the localization results for abnormal event in UMN dataset. To detect block-level abnormality, $\beta = 0.3213$ is used, and each regional BoV contains 20 volumes.



(a) (b) (c)

Figure 3: Results of the scene partitioning in UMN dataset. (a), (b) and (c) are the scene partitioning results in 3 videos of the UMN dataset.



(a)

(b)

(c)

Figure 4: Localization results of abnormal events in the UMN dataset. Detection results in the frames are highlighted in red boxes. Right images are original frames. Left images are localization results of right images.

(a) (b) (c)

Figure 5: Sample frames of the GIST YOUTUBE dataset. (a) is 'building collapse' scenario. (b) is frame which is included to 'shootings' scenario. (c) is frame of 'gas explosion' scenario.

4.2 GIST Youtube Dataset

To evaluate the performance of the proposed method in practical scenes, we conducted experiments on the dataset which consists of videos that are downloaded from the Youtube. The dataset comprise of 3 videos

Table 1: Performance comparison for frame-level abnormal event detection in the UMN dataset.

Method	AUC
Pure optical flow	0.84
SFM	0.96
Proposed method	**0.9938**

for abnormal events such as a building collapses, shooting, gas explosion.[1] Figure 5 shows snapshots of each video for the normal events. Each video includes sufficient training data to build a pattern model of normal activities. The 'building collapse' video contains 805 frame, the 'shootings' video consists of 3,099 frames, and the 'gas explosion' video has 4,640 frames. All frames in videos was resized to the width of 320 pixels and height of 240 pixels.

The size of the spatio-temporal volume and the size of regional BoVs are equal to the previous experiments in Section 4.1. In 'Building collapse' video, $\alpha = 0.1$ is used to the threshold for the scene partitioning, $\alpha = 0.1$ is used for 'Shooting' video, and $\alpha = 0.095$ is used for 'Gas explosion' video. Figure 7 show results of scene partitioning of each video in the GIST Youtube dataset.

The detection results are shown in Fig. 6. The normal and abnormal results are annotated as red and greed colors in the indicated bars. Figure 8 shows the localization results when the proposed method detects the abnormal events within a frame.

5 CONCLUSION

Videos obtained from a surveillance system can contain a lot of diverse moving objects, and these moving objects can represent various activity patterns. Furthermore, according to the geometric conditions of cameras, the activity patterns can be distorted. Due to these reason, detection of abnormal event is one of the challenging issues in computer vision research fields.

In this paper, we proposed the novel method using scene partitioning for abnormal event detection in complex situations. To partition a scene, we divide given frames into 3D blocks, and we compute the dense optical flow to represent activity patterns of each block. In order to describe the characteristics of activity patterns of each block and remove redundant information, we computed the entropy and magnitude of HoF of each 3D block. Finally, we define regions through grouping blocks by measuring a similarity based on the entropy and magnitude,

[1] The dataset is available at http://mlv.gist.ac.kr/mlvarch_dataset/gist_anomaly_dataset

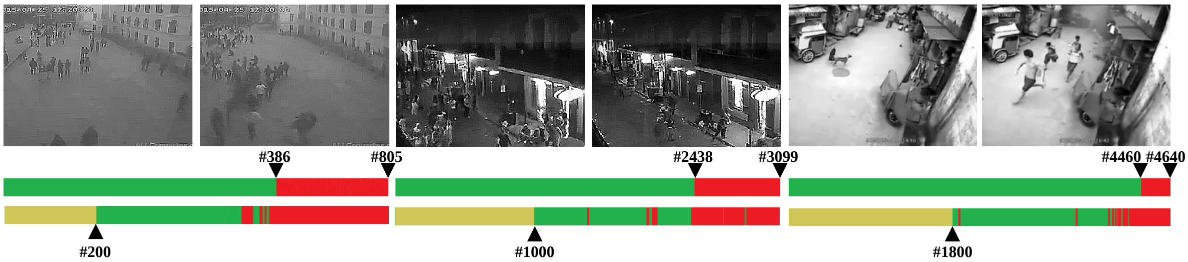


Figure 6: Qualitative results of the abnormal event detection for three videos of the GIST Youtube dataset.The top row represents snapshots of video in the dataset. At the bottom, the ground truth bar and the detection result bar show the abnormality of each frame in videos, where the green color denotes the normal frame, the red color describes the abnormal frame, and the yellow color corresponds to frames which are use to scene partitioning and training the regional BoVs.

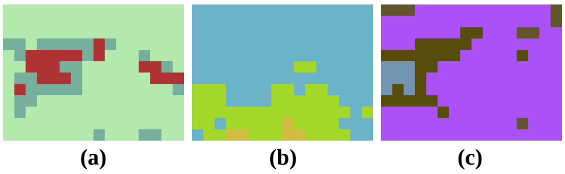


Figure 7: Results of the scene partitioning based on the regional activity pattern analysis in GIST Youtube dataset. (a) shows the result of scene partitioning of 'building collapse' video. (b) denotes the result of scene partitioning for 'shootings' video. (c) represent the scene partitioning result of 'gas explosion' video.

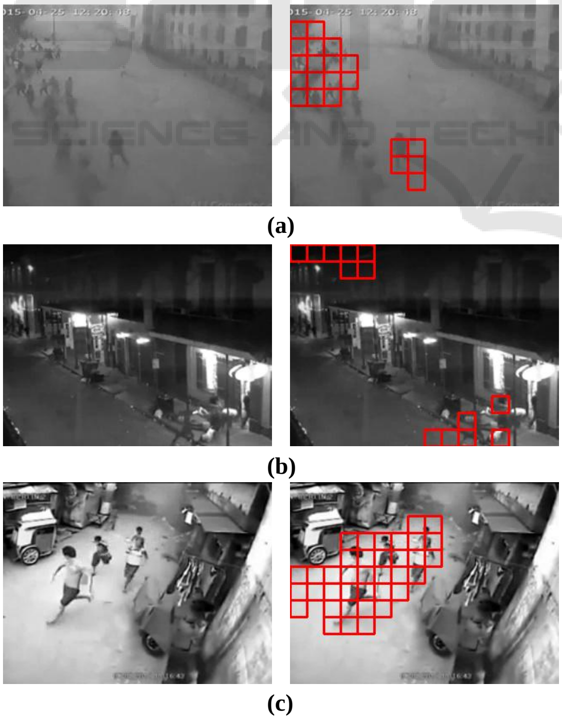


Figure 8: Localization results in the GIST Youtube dataset. Abnormal event location are highlighted in red boxes. Right images are original frames. Left images are localization results of right images.

and model the regional BoVs independently. In this manner, The proposed method can model the normal events of complex situations. The experimental results showed that the proposed abnormal event detection method can detects abnormal activities successfully in two datasets, and demonstrate that the method can provide an efficient way to detect the abnormal event in practical scenes.

The current version of proposed method has two major drawbacks. First, in order to perform scene partitioning precisely, the proposed method requires a sufficient number of observations for moving objects. When the moving objects are not observed, or the number of the observed objects is not sufficient, results of the scene partitioning and the normal event modelling can degrade performance of abnormal event detection. Second, the proposed method determines thresholds manually by users. The inaccurate setting of thresholds can be the main cause incorrect detection for abnormal events. However, these drawbacks are general problems appeared in most of abnormal event detection methodologies.

To this end, our future work is to improve the result of the scene partitioning by incorporating appearance information, and develop an automated method for determining suitable threshold values.

ACKNOWLEDGEMENTS

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.B0101-15-0525,Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis) and Center for Integrated Smart Sensors as Global Frontier (CISS-2013M3A6A6073718).

REFERENCES

Alvarez, L., Weickert, J., and Sánchez, J. (2000). Reliable estimation of dense optical flow fields with large displacements. *International Journal of Computer Vision*, 39(1):41–56.

Andrade, E. L., Blunsden, S., and Fisher, R. B. (2006a). Hidden markov models for optical flow analysis in crowds. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 460–463. IEEE.

Andrade, E. L., Blunsden, S., and Fisher, R. B. (2006b). Modelling crowd scenes for event detection. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 175–178. IEEE.

Bayona, Á., SanMiguel, J. C., and Martínez, J. M. (2009). Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 25–30. IEEE.

Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE.

Cong, Y., Yuan, J., and Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3449–3456. IEEE.

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

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Goldberger, J., Gordon, S., and Greenspan, H. (2003). An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 487–493. IEEE.

Javed, O. and Shah, M. (2002). Tracking and object classification for automated surveillance. In *Computer VisionECCV 2002*, pages 343–357. Springer.

Klaser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association.

Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., and Shafer, S. (2000). Multi-camera multi-person tracking for easyliving. In *Visual Surveillance, 2000. Proceedings. Third IEEE International Workshop on*, pages 3–10. IEEE.

Li, W., Mahadevan, V., and Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):18–32.

Liem, M. C. and Gavrila, D. M. (2014). Joint multi-person detection and tracking from overlapping cameras. *Computer Vision and Image Understanding*, 128:36–50.

Lucas, B. D., Kanade, T., et al. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679.

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

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

Moon, T. K. (1996). The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60.

Pan, J., Fan, Q., and Pankanti, S. (2011). Robust abandoned object detection using region-level analysis. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 3597–3600. IEEE.

Roshtkhari, M. J. and Levine, M. D. (2013). An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer vision and image understanding*, 117(10):1436–1452.

Sato, K. and Aggarwal, J. (2001). Tracking and recognizing two-person interactions in outdoor image sequences. In *Multi-Object Tracking, 2001. Proceedings. 2001 IEEE Workshop on*, pages 87–94. IEEE.

Tian, Y., Feris, R. S., Liu, H., Hampapur, A., and Sun, M.-T. (2011). Robust detection of abandoned and removed objects in complex surveillance videos. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(5):565–576.

Wang, B., Ye, M., Li, X., Zhao, F., and Ding, J. (2012). Abnormal crowd behavior detection using high-frequency and spatio-temporal features. *Machine Vision and Applications*, 23(3):501–511.

Wang, S. and Miao, Z. (2010). Anomaly detection in crowd scene using historical information. In *Intelligent Signal Processing and Communication Systems (ISPACS), 2010 International Symposium on*, pages 1–4. IEEE.

Zhou, S., Shen, W., Zeng, D., and Zhang, Z. (2015). Unusual event detection in crowded scenes by trajectory analysis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1300–1304. IEEE.