Self-Learning pLSA Model for Abnormal Behavior Detection in Crowded Scenes

SUMMARY
Abnormal behavior detection is now a widely concerned research field, especially for crowded scenes. However, most traditional unsupervised approaches often suffered from the problem when the normal events in the scenario with large visual variety. This paper proposes a self-learning probabilistic Latent Semantic Analysis, which aims at taking full advantage of the high-level abnormal information to solve problems. We select the informative observations to construct the “reference events” from the training sets as a high-level guidance cue. Specifically, the training set is randomly divided into two separate subsets. One is used to learn this model, which is defined as the initialization sequence of “reference events”. The other aims to update this model and the infrequent samples are chosen into the “reference events”. Finally, we define anomalies using events that are least similar to “reference events”. The experimental result demonstrates that the proposed model can detect anomalies accurately and robustly in the real-world crowd environment.

key words: video surveillance, abnormal behavior detection, action transfer rules, pLSA

1. Introduction
Abnormal detection in video surveillance can significantly improve public safety and security [1], [2]. However, it is a difficult problem especially in real life due to the diversity and complexity of behaviors. Most previous methods are unsupervised and based on an assumption that abnormal behaviors are rare and minor in contrast to the dominant normal behavior. However, the unsupervised approaches often suffered from the problem when the normal events in the scenario with large visual variety, since they did not consider any abnormal information. Unlike the unsupervised approach, the supervised approach can achieve high accuracy. However, few approaches are supervised, because accurately labeling all samples is an exhaustive work when meeting the large dataset.

From the strengths and limitations of two methods, this paper proposes a self-learning probabilistic Latent Semantic Analysis (self-learning pLSA) model, which is an extension of pLSA model. This model discovers the behavior distributions through the topic model with the aid of the high-level abnormal information. Form the observation its current advances, the topic model has been applied into the abnormal detection issue, such as [3]–[5]. One drawback of the traditional approach based on the topic model is overlooking any high-level semantic information. This may be appropriate for text but not for sensory data with large variety in appearance. To this end, we select the informative observations to construct the “reference events” from the training sets as a high-level guidance cue. The training set is randomly divided into two separate subsets, i.e. the model learning data and model updating data, respectively. The model learning data is first used as the initialization sequence of “reference events” to learn the action transfer rules (ATRs) and behavior distributions. And then the behavior distribution of the model updating data is calculated with learn ATRs kept fixed. Their ATRs and behavior distributions are updated through the abnormal value which is acquired by comparing its behavior distributions with respect to all of the reference events. Anomalies are identified as these events that display a significant difference of occurring based on “reference events”. The experimental result demonstrates that the proposed model can detect anomalies accurately and robustly in the real-world crowd environment.

2. Self-Learning pLSA Model for Abnormal Behavior Detection in Crowded Scenes
In this paper, we make use of the actions transition of frame pairs for representation that, equipped with a novel generative model, performs abnormal detection task. Figure 1 illustrates the flowchart of our framework. At the first level, each frame pair is represented as an action transition matrix. In the subsequent level, a novel self-learning pLSA model is introduced to construct the behavior distribution through the learning and updating process.
the model learns the ATRs and behavior distribution using the model learning data. In the updating process, each frame pair of the model updating data is first used to calculate the behavior distribution \( p(z | f_{t1}, f_{t2}) \) with the learnt ATR \( p(a_j, a_i | z) \) kept fixed. The two parameters are then updated through the abnormal value which is acquired by comparing their behavior distributions with respect to all of the reference events. Finally, the behavior distribution of test observation is estimated with the learnt ATRs kept fixed. Anomalies are defined as the observations having very low abnormal values.

2.1 Action Transition Matrix

The visual features are extracted from a set of non-overlapping patches of fixed size to cover entire video frames. There uses optical flow proposed by [6] to represent frames, as it has been proven to be a useful low-level feature in many computer vision applications. Generally, optical flow corresponds to the instantaneous velocity of the moving pixels on objects, calculated by exploring the variation of pixels between consecutive frames. Optical flow statistical histograms are constructed separately according to their spatial location both in direction and magnitude.

Based on the above patch descriptors, it is necessary to construct visual words which can efficiently represent the behavior content. Since patches from different images may share similar visual properties in terms of the local descriptors, we use K-means to perform the clustering on all patches. Each cluster is defined as an action, and we obtain a collection of discrete actions \( A = \{a_i\}_{i=1}^{n} \), in which \( n \) is the number of actions. Thereafter, each local descriptor is further quantized into one of the clusters according to the nearest neighbor rule. Thus a given frame is represented a histogram \( H = [h_i]_{i=1}^{n} \), which the bin \( h_i \) is the value of the \( i \)th actions. This paper models each frame pair as the action transition matrix, which is defined as \( ATM = H_{frame_1} H_{frame_2} = (\text{act} \tan \frac{h_i}{m}), H_{frame_1} = [h_i]_{j=1}^{m}, H_{frame_2} = [h_i]_{j=1}^{m}. \) Intuitively, the \( j \)th row of the matrix measures the correlation between all bins values from frame \( p \) and \( j \)th bin value of frame \( q \).

2.2 The Learning Process of Self-Learning pLSA Model

Self-learning pLSA model is the extension of pLSA model [3]. pLSA model has received considerable interest in the text analysis community as a tool to model documents as a mixture of several semantic-but a-prior unknown, and hence “latent topics”. This model catches our attention as the anomaly is determined by a combination of multiple behavior distributions, the behavior of which is a hidden variable.

Given a set of frame pairs \( F = \{f_1, \ldots, f_p\} \) of the model learning data, each is described in a action transition matrix \( ATM \), which is generated by latent behavior \( Z = \{z_1, \ldots, z_M\} \). In the self-learning pLSA model, the conditional probability of the pair-wise of actions \( p(a_j, a_i | f) \) as:

\[
p(a_j, a_i | f) = \sum_{\Delta \in \Delta} (a_j, a_i | \Delta) p(\Delta | f)
\]

where \( p(a_j, a_i | \Delta) \) is the probability of pair-wise of actions \( a_j \) and \( a_i \) when the behavior is \( \Delta \). \( p(\Delta | f) \) denotes behavior distributions for frame pair.

The two parameters can be estimated with an Expectation-Maximization (EM) algorithm [3] by maximizing the log-likelihood function:

\[
L = \sum_{a_j, a_i \in A} \frac{n(f, a_j, a_i)}{\frac{1}{2}(1 - \cos(\pi \tau))}
\]

where \( p(f, a_j, a_i) = p(f)p(a_j, a_i | f) \), and \( n(f, a_j, a_i) \) is the term frequency. \( p(f) \) is used to denote the selection probability for the frame \( f \), which is proportional to its length. \( n(f, a_j, a_i) \) denotes the number of times the pair-wise of actions \( (a_j, a_i) \) occurred in frame \( f \).

2.3 The Updating Process of Self-Learning pLSA Model

In the updating process, each frame pair of the model updating data is first calculated the behavior distribution \( P(z | f_{t1}, f_{t2}) \) with the learnt ATRs \( P(a_j, a_i | z) \) kept fixed in EM manner according to Eq. (2). In E-step, the posterior probabilities for \( P(z | f_{t1}, f_{t2}, a_j, a_i) \) are calculated, and in M-step, \( P(a_j, a_i | z) \) and \( P(z | f_{t1}, f_{t2}) \) are updated. And then the ATRs and behavior distributions are updated through the abnormal value.

We calculate the abnormal values by comparing their behavior distributions with respect to all of the reference events. Assuming \( Q \) samples in the reference events, their behavior distributions are defined as \( P(z | f_1), \ldots, P(z | f_Q) \). For observation \( f_1 \), a similarity vector \( S_1 \) is generated as:

\[
S_1 = (C(P(z | f_1), P(z | f_1), \ldots, C(P(z | f_1, P(z | f_Q)))) \quad (3)
\]

where \( C \) is the histogram correlation.

The abnormal value \( \sigma_1 \) is defined as a weighted average of \( S_1 \) as: \( \sigma_1 = \omega^T S_1 \), where \( \omega \) is the exponentially decaying weights \( \omega = \frac{1}{\lambda} e^{-\lambda \Delta t}, e^{-\lambda \Delta t}, \ldots, e^{-\lambda \Delta t} \), where \( \lambda \) is the decay constant.

Finally, the estimated abnormal value is used to update the ATRs and behavior distribution via running EM in a similar manner by maximizing the \( L \) function:

\[
L = \sum_{a_j, a_i \in A} \frac{n(f_{t1}, f_{t2}, a_j, a_i)}{\frac{1}{2}(1 - \cos(\pi \tau))}
\]

where \( p(f_{t1}, f_{t2}, a_j, a_i) = p(f_{t1}, f_{t2}) p(a_j, a_i | f_{t1}, f_{t2}), n(f_{t1}, f_{t2}, a_j, a_i) \) is the term frequency, containing the action occurrences for every frame pair.

The parameter \( \rho \) is an adjusting parameter, whose range is between 0 and 1. \( \rho \) is set as 1 in the learning process, which means that the adjust factor has no influence on the
maximum log-likelihood. In the updating process, the observations are used to adopt \( p(a_j, a_i | z) \) if cosine value of temporal stability \( \sigma_{ij} \) nearly \(-1\), otherwise that observations are treated as redundancy. Meanwhile, the informative observations are added to the "reference events".

Considering a test frame, its abnormal value is calculated according to Eq. (3). A large abnormal value indicates a normal frame and the abnormal value is small for abnormal frames. In experiment, the anomaly is defined when the abnormal value \( \sigma_t < 0.5 \).

3. Experiment Results

In this section, we introduce a new dataset to evaluation of the effectiveness of our method and compare its performance with state-of-the-art methods on UMN [7] and UCSD [8]. We construct a new railway station dataset (RS), which is a collection of anomalies in several passenger railway stations. The detection of abnormal events in a railway station is of great importance to the station management and to the public safety, as there are many crowd scenarios, the abnormal events of which may lead to panics and further to fatal consequences. The RS dataset contains 10065 videos from 190 stations recorded in the four areas, i.e. the entrance and exit gate (EEG), the exit area (EA), the waiting room (WR) and the platform (PF). There are 1961 videos for the entrance and exit gate area. Among the videos, there are 166 videos considered as abnormal events, which show people walking in the opposite direction. For the exit area, there are 943 videos in total, among which are 119 abnormal events showing people walking in the wrong direction or loitering near the exit. There are 2354 videos for the waiting room and the 368 anomalous events, which include the sudden gathering and dispersing, the sudden stop of walking, the running and etc. There are 2873 videos for the platform area, 580 of which indicate invalid behaviors such like standing beyond the yellow line.

We first train this model based on the RS dataset. And then this model is tested using ten videos collected real-time from the Beijing railway station and Changsha railway station in the different areas, respectively. The experimental results in terms of the area under the ROC (AUC) are shown in Table 1. Looking closely at the Table 1 from the railway station perspective, we can find out that the accuracies are generally higher in the Changsha station than in the Beijing station. We believe the difference is related to the resolution and layout of the video from these two stations. The analysis of Table 1 by area is as follows: the best performance is achieved in the entrance and exit gate. The main reason is that the abnormal behaviors in this area are relatively simple and highly discriminative. It is usually difficult to accurately detect anomalies in the waiting room, comparing with other three areas. Because many behaviors are occluded. Also, due to the lack of enough illumination, many behaviors are not accurately estimate.

Figure 2 gives the performance comparison between the proposed self-learning pLSA with traditional pLSA in the abnormal detection. The pLSA is a generative model which aims to find latent topics from visual words given a set of frames. It can be seen that the pLSA model can not perform well. The main reason is that the pLSA model learns the topics by grouping low-level features without considering any high-level abnormal information. In contrast to concentrate on the low-level features only, the proposed self-learning pLSA model designs also focus on the high-level semantic information.

We subsequently investigate how the classification performance is affected by the number of behavior. Figure 3 shows the performance variation for the five differently sized behavior variables \{10, 30, 50, 80, 100\}. It can be seen that the performance achieves the best when the number of rules equals to 50. The performance progressively improves until behavior size is 50, and then drops off slightly. It indicates that if the number is too small, it will be difficult to find the action transfer rules and abnormal events may be grouped with the normal ones. If the number of behaviors is too large, some normal activities will be matched to the abnormal, which brings down the detection performance.

Figure 4 compares the the proposed method against recently results in the UMN dataset. This data set is a public available data set of anomalous events, which is used to compare the performance of the most methods. In addition, the abnormal behavior in the data set is a common occurrence in public places. The UMN dataset contains three different escape scenes (named as lawn, indoor, and plaza, respectively), in which pedestrians initially walk randomly and exhibit a sudden evacuation in different directions in the end. We use the ROC curves as evaluation

| Table 1 | The performance of the proposed model on RS dataset |
|---------|-----------------------------------------------|
|         | WR   | EEG  | EA   | PF   |
| STA.Beijing| 88.3% | 94.5% | 91.8% | 92.7% |
| STA.Changsha| 90.2% | 95.8% | 92.6% | 93.1% |

Fig. 2 The performance comparison with the traditional pLSA model

Fig. 3 The performance over different behavior sizes
measurement. As the global anomalies can be easily detected in the UMN dataset, our method also achieves perfect performance, which reaches 99.7% under the AUC criterion.

We also carry out experiments on another real-world video dataset, the UCSD Ped 1 dataset. This dataset is a challenging dataset due to low resolution, many occlusions, different kinds of anomalies, and co-occurring in the same frame. Only the pedestrian walking on the walkway is treated as normal behavior, while abnormal behaviors have various kinds, such as cycling, skating, and vehicle crossing. We use the Equal Error Rate (EER) as reliable criterion, which is the ratio of misclassified frames at the point of FPR = TPR on the curve. The results in Table 2 demonstrate that the proposed method outperforms other approaches. We attribute this to the fact that it accounts for abnormal information in the scene and hence, it is capable of learning complicated behaviors. Although adding abnormal information increases the computational complexity of the self-learning pLSA algorithm when compared to local approaches, it is still fast enough for real-time abnormality detection.

4. Conclusion

This paper proposes a self-learning pLSA model to detect anomalies in crowded scenes. We have observed that the unsupervised method can not well solve the situation when the normal events in the scenario with large visual variety. In view of this, we try to address problem by taking full advantage of the high-level abnormal information. In future work, we plan to incorporate other cues into consideration to improve the detection performance.

Acknowledgments

This work was support by the fund of Winter Olympics of science and technology (2020YFF0304102).

References

[1] S. Xian and Z. Songhao, “Weak supervised learning based abnormal behavior detection,” Proc. ICPR., pp.1580–1585, 2018.
[2] L. WeiXin, V. Mahadevan, and N. Vasconcelos, “Anomaly detection and localization in crowded scenes,” IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no.1, pp.18–32, 2014.
[3] J. Varadarajan and J.-M. Odobez, “Topic Models for Scene Analysis and Abnormality Detection,” Proc. ICCV Workshop on Visual Surveillance, pp.1338–1345, 2009.
[4] R. Kaviani, P. Ahmadi, and I. Gholampour, “Incorporating fully sparse topic models for abnormality detection in traffic videos,” Proc. ICCKE, 2014.
[5] J. Li, S. Gong, and T. Xiang, “Global Behaviour Inference using Probabilistic Latent Semantic Analysis,” Proc. BMVC, 2008.
[6] Y. Zhang, H. Lu, L. Zhang, and X. Ruan, “Combining motion and appearance cues for anomaly detection,” Pattern Recognition, vol.3, no.51, pp.443–452, 2016.
[7] Unusual crowd activity dataset of University of Minnesota, available from http://mha.cs.umn.edu/movies/crowd-activity all.avi.
[8] V. Mahadevan, W. Li, V. Bhaldia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” Proc. CVPR, pp.1975–1981, 2010.
[9] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” Proc. CVPR, pp.935–942, 2009.
[10] S. Wu, B. Moore, and M. Shah, “Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes,” Proc. CVPR, pp.2054–2060, 2010.
[11] Y. Cong, J. Yuan, and J. Liu, “Sparse reconstruction cost for abnormal event detection,” Proc. CVPR, pp.3449–3456, 2011.
[12] V. Saligrama and Z. Chen, “Video anomaly detection based on local statistical aggregates,” Proc. CVPR, pp.2112–2119, 2012.
[13] Z. Shifu, S. Wei, and Z. Dan, “Spatial-temporal convolutional neural networks for anomaly detection and location in crowded scenes,” Signal Processing: Image Communication, vol.44, no.1, pp.358–368, 2016.