Study on Anomaly Detection in Crowd Scene

Jun zhang [1, a,*], Yunxia Chu [2, b]

[1] School of Computer Science, Shijiazhuang University, Shijiazhuang 050035, China

[2] School of Fine Arts, Shijiazhuang University, Shijiazhuang 050035, China

[a]email: zhang72jun@163.com, [b]chuyunxia1980@sina.com, [*]corresponding author

Keywords: Crowd Scene, Anomaly detection, Bag-of-words, Probabilistic Latent Semantic Analysis(PLSA), Interest Points

Abstract. Anomaly detection technology in crowd scene is very important in public place. Crowd detection differs from pedestrian detection which we assume no individual pedestrian can be properly segmented in the image. We propose a scheme which the scen can be treated the crowd motion patterns as the spatial-temporal domain. In the classification stage, we divide whole frame into small blocks, and motion pattern in each block is encoded by the distribution of motion bags in it. PLSA classifier is proposed to infer classification of crowed detection, and we classify motion pattern into normal or abnormal group according to the deviation between motion pattern and train model. The comprehensive implementation can detect crowd in real-time. This paper presents an approach to automatically detect abnormal behavior in crowd scene with Interest points to represent moving objects to generate word of bags, which are used to describe crowed moriment results show that the speed of detection has been greatly improved using our approach.

Introduction

Visual analysis has a potential to be used for recognition, and it is one of the hottest but most difficult subjects in computer vision. With the public safety becomes a very important issue in public place, such as subway station, airport, shopping mall and square, deployment of intelligence surveillance system is more and more common. This system is designed to detect abnormal behavior automatically, so people can take action as soon as possible to prevent dangerous events. However, because of the high crowd density in such scene, although a lot of works have been done for this goal, it is still a very challenging thing. There may be thousands of behaviors in crowd scene, it is almost impossible to define every of them, and lack of abnormal video data makes this problem even worse [1].

In this paper, we present an approach to automatically detect abnormal behavior in crowd scene with Interest points as feature points to represent moving objects and tracked by optical flow technique to generate motion vectors, which are used to describe crowed motion feature. We divide whole frame into small blocks, and just extract motion pattern to represent the motion in each block. We use Interest points as feature points to represent moving objects and track these points by optical flow technique, through this way, we get motion vectors to describe motion, which include velocity and direction information. We do not define very behavior specifically, we just define two behavior groups: normal behavior and abnormal behavior. We cluster similar normal motion patters into normal pattern models in an on-line method, then we compute deviation between new coming patterns and trained models to do classification. In the classification stage, we divide whole frame into small blocks, and motion pattern in each block is encoded by the distribution of motion vectors in it. Deep Learning classifier is proposed to infer classification of crowed detection, and we classify motion pattern into normal or abnormal group according to the deviation between motion pattern and trained model. We propose a scheme that looks at the motion patterns of crowd in the spatial-temporal domain and give an efficient implementation that can detect crowd in real-time.

RELATED WORK

Nowadays, there are a lot of researches related to anomaly detection. According to different criterion, all these works can be classified in different categories. On the basis of whether analysis every individual’s activity, there are two categories: tracking stage and holistic stage. Tracking stage needs to segment and track every individual in scene, then analysis behavior by compute the trajectory, speed and direction of individual [2]. Such approach is subject to impaction of occlusion, so it is not suitable for high density crowd. On the contrary, Holistic stage consider crowd as a entirety. It does not try to detect and track individual, just extract some features (corner, gradient, optical flow) to represent motion [1-4]. For example, Jaechul Kim et al. [5] approach divide whole frame into small cuboid and compute optical flow in every cuboid. They generate activity pattern using MPPCA, then do Bayesian inference on MRF. Such approach needn’t segment every individual so it is more robust for crowd scene.

Researches on group behaviors can be mainly divided into three categories. The first is the traditional object-based approaches which consider the group as a collection of individuals [6]. This kind of methods analyzes the crowd behaviors through individuals. In simple situations which have a small number of moving objects, this kind of methods can achieve good results. However, in the complex scene, there exist severe occlusions. It is almost impossible for object segmentation, tracking and behavior recognition. The computational cost will also be greatly influenced by the number of objects (e.g., people). Another category of methods focuses on analyzing the entire video frame and extracting the subject specific information [7].

Our main contributions can be summarized as follows. 1) An extraction method to the size-adapted spatio-temporal cuboid is proposed. The positions and total number of the cuboids can be determined automatically. 2) Compared with the previous methods, the detection accuracy can be significantly improved.

The organization of this paper is as follows. Related work is reviewed in next section. Section 3 describes the motion pattern descriptor. In section 4 we introduce how to cluster patterns into models. In section 5 we demonstrate experiment results on different database.

MOTION DESCRIPTOR

Interest Point Methods. Now a popular approach in object recognition is based on spatial-temporal interest points and local feature descriptors. All of the local descriptors are usually vector-quantized to obtain a finite set of visual words before they are fed into any classification algorithms [8]. Laptev et al. [7] propose a spatial-temporal interest point operator, which detect local structures in space-time, and the local structures have large variations both in space and time. Niebles et al. [8] combine shape information with local appearance features by building a hierarchical model, which can be characterized as a constellation of bag-of-features. Local descriptors extracted from space-time interest points have also been shown to work well on videos with complex scenes.

Assuming a stationary camera or a process that can account for camera motion, separable linear filters are applied to the video to obtain the response function as follows:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel, only applied along the spatial dimensions (x, y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as $h_{ev} = (t; \tau, \omega) = -\cos(2\pi t\omega) \times e^{-t^2/\tau^2}$ and $h_{od} = (t; \tau, \omega) = -\sin(2\pi t\omega) \times e^{-t^2/\tau^2}$. The two parameters σ and τ correspond to the spatial and temporal scales of the detector respectively. In all the following experiment, we assume $\omega = 4/\tau$, thus reduce the two number of parameters in the response function $\mathbb{R}$. To handle multiple scales, we must run the detector over a set of spatial and temporal scales.

They [8] note any region with spatially distinguishing characteristics undergoing a complex motion can induce a strong response. However, regions undergoing pure translate motion or without spatially distinguishing features will not induce a strong response. The space-time interest points are extracted around the local maxima of the response function. Each patch contains the volume that contributes to the response function.

Topic Models for Visual Recognition. Recently, Bag-of-Words [6,9] models have drawn more attention recently in the field of object recognition. The Bag-of-Words model is originally proposed for analyzing text documents, where a document is represented as a histogram over word counts. Generative topic models are then applied to this Bag-of-Words representation, and the topics of the document are denoted as latent variables in these models [9].

Despite the considerable achievements in the field accomplished in recent years, there are still some challenges to be overcome. The optimal abnormal behavior detecting should allow the detection of suspicious events with a minimal description and perform the detection without assumptive scenario [7], and the most important can recognize more abnormal behavior using the same library. For this goal to be achieved, we develop a double-layer Bag-of-Words model, which is described and evaluated in this paper.

Feature Representation. There are several choices in the selection of good features to describe moving objects. In general, there are three popular types of features: static features based on edges, dynamic features based on optical flow measurements, and spatial-temporal features obtained from local video patches. In particular, spatial-temporal interest points have turned out to be useful in the human motion categorization task, providing a rich description and powerful representation [1].

CLUSTER MOTION PATTERNS

As the left of Fig.1 illustrates, we represent each video sequence as a collection of spatial-temporal words by extracting space-time interest points. Among the available interest point detectors for video data, the interest points obtained using the generalized space-time corner detector is too sparse to characterize many complex videos. The separable linear filter method in [1] is adopted, since it generally produces a high number of detections. In the following, we provide a brief review of the detector proposed in [1].

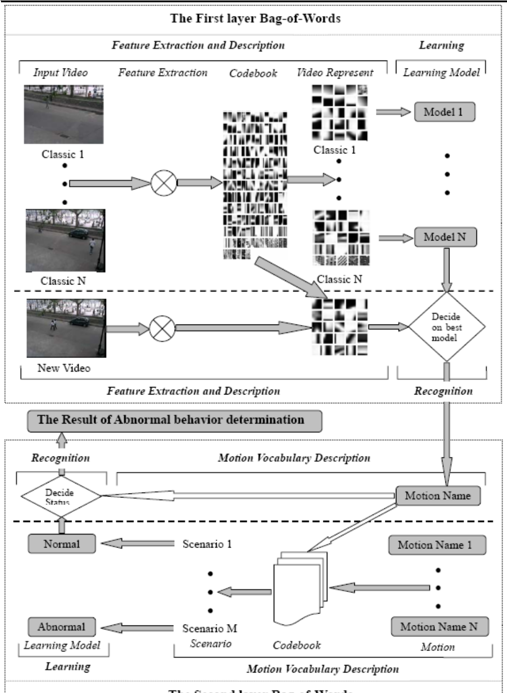


Fig. 1 Flowchart of semi-supervised abnormal action recognition in crowed scence

To obtain a descriptor for each spatial-temporal cube, we calculate its brightness gradients in x, y andt three directions. The spatial-temporal cube is then smoothed at different scales before the image gradients are computed. The computed gradients are concatenated to form a vector. The size of the vector is equal to the number of pixels in the cube times the number of smoothing scales times the number of gradients directions. This descriptor is then projected to a lower dimensional space using the principal component analysis (PCA) dimensionality reduction technique. In [1], different descriptors have been used, such as normalized pixel values, brightness gradient and windowed optical flow. Both the gradient descriptor and the optical flow descriptor are equally effective in describing the motion information [5].

This vocabulary (or codebook) is constructed by clustering using the k-means algorithm and Euclidean distance as the clustering metric. The center of each resulting cluster is defined to be a spatial-temporal word (or codeword). Thus, each detected interest point can be assigned a unique cluster membership, i.e., a spatial-temporal word, such that a video can be represented as a collection of spatial-temporal words from the codebook. The effect of the codebook size is explored in our experiments. In the text codebook, the latent topic models pLSA and LDA rely on the existence of a finite vocabulary of scenario-motion number with size M.

Learning and Recognizing the Anomalous by h-pLSA. LSA is a theory and a method for extracting and representing the contextual usage meaning of words by statistical computations applied to a large corpus of text. In the following, we provide a brief review of learning and recognizing the action model by pLSA, which proposed in [9], and we will describe our method for abnormal behavior learning and recognition in different scenarios using h-pLSA.

Learning and Recognizing the Action Model by pLSA. Suppose there is $N(j=1,\cdots,N)$ video sequences containing spatial-temporal words from a vocabulary of size $V(1,\cdots,V)$. The corpus of videos is summarized in a $V\times N$ co-occurrence table $\bar{N}$, where $m(w_i,d_j)$ stores the number of occurrences of a spatial-temporal word w_i in a video d_j. In addition, there is a latent topic variable z_k associated with each occurrence of a spatial-temporal word w_i in a video d_j. Each topic corresponds to an action category, such as walking, running, etc.

The joint probability $P(w_i,d_j,z_k)$ is assumed to have the form of the graphical model shown in horizontal direction of Fig. 2. Nodes are random variables. Shaded ones are observed and unshaded ones are unobserved. The plates indicate repetitions. In the context of human action categorization, d represents video sequences, z are action categories, and ware spatial-temporal words. The parameters of this model are learnt in an unsupervised manner using an EM procedure [9].

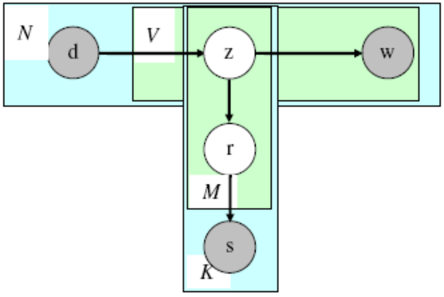


Fig 2 The hybrid probabilistic Latent Semantic Analysis (h-pLSA)

Given that the observation pairs (d_j,w_i) are assumed to be generated independently, we can marginalize over topics z_k to obtain the conditional probability $P(w_i\,|\,d_j)$:

$$P(d_j,w_i)=P(d_j)P(w_i\,|\,d_j) \qquad (2)$$

where $P(z_k\,|\,d_j)$ is the probability of topic z_k occurring in a video d_j , and $P(w_i\,|\,z_k)$ is the probability of spatial-temporal word w_i occurring in a particular action category z_k. There is the total number of K latent topics, which is the number of action categories.

They then fit the model by determining the action category histograms $P(w_i\,|\,z_k)$ and the mixture coefficients $P(z_k\,|\,d_j)$. In order to determine the model that gives the highest probability to the spatial-temporal words that appear in the corpus, a maximum likelihood estimation of the parameters is obtained by maximizing the following objective function using an expectation-maximization (EM) algorithm:

$$\prod_{i=1}^{V}\prod_{j=1}^{N}P(w_i \mid d_j)^{m(w_i,d_j)} \tag{3}$$

where $P(w_i \mid d_j)$ is given by (3).

The algorithm has latent the action category models, the goal is to categorize new video sequences. The action-category-specific video-word-distributions $P(w \mid z)$ from a different set of training sequences have been obtained. When a new video is given, the unseen video is projected on the simplex spanned by the learnt $P(w \mid z)$. The next step is finding the mixing coefficients $P(z_k \mid d_{test})$ such that the KL divergence between the measured empirical distribution $\tilde{P}(w \mid d_{test})$ and $P(w \mid d_{test}) = \sum_{k=1}^{K} P(z_k \mid d_{test}) P(w \mid z_k)$ is minimized [9]. Similar to the learning stage, an EM algorithm to find the solution is applied. Thus, a categorization decision is made by selecting the action category that best explains the observation, that is:

$$Action\ Category = \arg\ \max_{k} P(z_k, d_{test}) \tag{4}$$

Also in the text with word recognition classification for learning and classification of the scene corresponding action text word pLSA method, graph model probability $P(z_i, s_j, r_k)$ is represented by Figure 3 vertical direction. The corresponding variable is shown in table 1.

Tab.1 The relationship between variable and variable

	Video		Text	
	Express	Explain	Express	Explain
Learning video, text number	v	Video frames	M	Scene of action number
Recognition of video, text number	N	Video frames	K	Scene of action number
Expressed sequence	d	A new video sequence	s	A new scene
Video, text set	w	Video collection	z	The set of action types
Subject headings	z	Movement types	r	Abnormal behavior

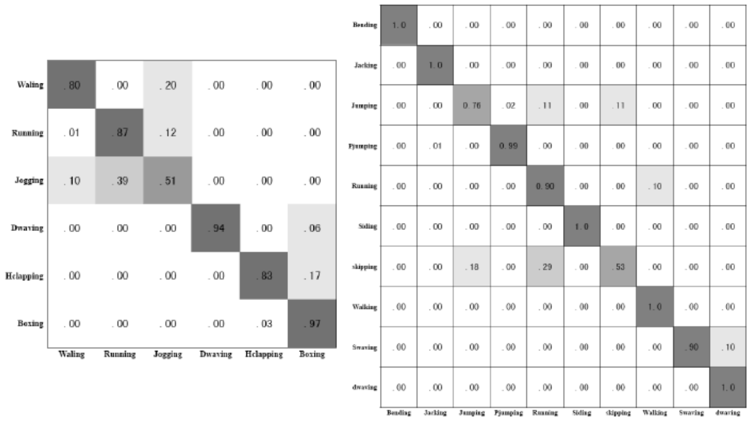
Recognizing the Anomalous by h-pLSA. In Fig. 3 we assume that the number of action categories in the horizontal direction is equal to the number of motion categories in the vertical direction, i.e. K in the subsection Learning and Recognizing the Action Model by pLSA is equal to K in the subsection Learning and Recognizing the Text by pLSA. Finally we draw the conclusion:

$$\mathrm{Re}\,sult = \arg\ \max_{m} P\left(r_m, \frac{P(w_i \mid z_k) P(z_k \mid d_j)}{\sum_{l=1}^{K} P(w_i \mid z_l) P(z_l \mid d_j)}\right) \tag{5}$$

Conclusion

In order to test the feasibility of our approach to abnormal action detection, we experiment with a set of real-time videos by the IVSS in five different scenarios, which are meeting room, street, beside the car, lobby, and laboratory. In the process of establishing the video codebook, we make use of our algorithm on two datasets: KTH human motion dataset [4], and Weizmann human action dataset [2]. The KTH human motion dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in different scenarios of indoor and outdoors with scale variation. The Weizmann human action dataset contains 83 video sequences showing nine different people, each performing nine different actions. We also test the effect of the number of video combination codewords on recognition accuracy on both two datasets, as illustrated in Fig. 4 It shows some dependency of the recognition accurace on the size of the codebook.

The result of our method not only can directly recognize the thirteen actions in different scenario, but also can identify the undirect definition actions, such as fighting, theft, kicking cars, picklock. Representative abnormal frames from each scenario are shown in Fig. 3. Note that when the button in the bottom of the picture turns red, there is an abnormal action appearing in our system. It is difficult to compare different abnormal action recognition systems. Each system has been designed for a certain purpose. All of the tests indicate that the double-layer Bag-of -Words method can be a choice for solving the abnormal recognition.



	Waling	Running	Jogging	Dwaving	Hclapping	Boxing
Waling	.80	.00	.20	.00	.00	.00
Running	.01	.87	.12	.00	.00	.00
Jogging	.10	.39	.51	.00	.00	.00
Dwaving	.00	.00	.00	.94	.00	.06
Hclapping	.00	.00	.00	.00	.83	.17
Boxing	.00	.00	.00	.00	.03	.97

Confusion matrix for the KTH dataset using combination codewords

	Bending	Jacking	Jumping	Pjumping	Running	Siding	skipping	Walking	Swaving	dwaving
Bending	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
Jacking	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
Jumping	.00	.00	0.76	.02	.11	.00	.11	.00	.00	.00
Pjumping	.00	.01	.00	0.99	.00	.00	.00	.00	.00	.00
Running	.00	.00	.00	.00	0.90	.00	.00	.10	.00	.00
Siding	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00
skipping	.00	.00	.18	.00	.29	.00	.53	.00	.00	.00
Walking	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00
Swaving	.00	.00	.00	.00	.00	.00	.00	.00	.90	.10
dwaving	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0

Confusion matrix for the Weizmann dataset using combination codewords

Fig. 3: Confusion matrix for the KTH and Weizmann dataset using combination codewords

Acknowledgments

The authors gratefully acknowledge the financial supports of the support of science and technology of Hebei Province (Grant NO. 13211717), Science and Technology Research Projects for talent cultivation engineering of Hebei Province (Grant NO. A2013003021), and Specialized Research Fund for the Doctoral Program of Shijiazhuang University (Grant NO. 12BS008).

Reference

[1] Carolina Garate, Piotr Bilinski, Francois Bremond, Crowd Event Recognition Using HOG Tracker, IEEE International Workshop on Performance Evaluation of Tracking and Surveillance [C] , 2009.

[2] N. Ihaddadene and C. Djeraba, Real-time crowd motion analysis, IEEE Intl Conf. on Pattern Recognition [C], Dec. 2008, pp. 14.

[3] Louis Krate, Ko Nishino, Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models, Computer Vision and Pattern Recognition [J], 2009.

[4] A. Basharat, A. Gritai, and M. Shah, Learning Object Motion Patterns for Anomaly Detection and Improved Object Detection, Computer Vision and Pattern Recognition [J], 2008.

[5] Ramin Mehran, Alexis Oyama, Mubarak Shah, Abnormal Crowd Behavior Detection using Social Force Model, Computer Vision and Pattern Recognition [J], 2009.

[6] S.Saxena, F.Brmond, M.Thonnat, and R.Ma, Crowd behavior recognition for video surveillance, In 10th International Conference on Advanced Concepts for Intelligent Vision Systems [C], 2008

[7]ELGAMMAL A, DURAISWAMI R, HARWOOD D, et al. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance [C]. Proceedings of the IEEE, 2002, 90(7):1153-1163, Taormina, ITALY

[8]Weiming Hu, Tieniu Tan, Liang Wang and Steve Maybank, A survey on visual surveillance of object motion and behaviors. IEEE Trans. System, Man, And Cybernetics [J], 2004, 34(3): 334-352.

[9]J. C. Niebles, H.Wang, L. Fei-Fei. Unsupervised learning of human action categories using spatialtemporal words. International Journal of Computer Vision[J], 2008, 79(3): 299–318.