DETECTING CONTEXTUAL ANOMALIES OF CROWD MOTION IN SURVEILLANCE VIDEO

Fan Jiang, Ying Wu, Aggelos K. Katsaggelos

Electrical Engineering and Computer Science Department, Northwestern University
2145 Sheridan Rd, Evanston, IL 60208
{fji295, yingwu, aggk}@eecs.northwestern.edu

ABSTRACT

Many works have been proposed on detecting individual anomalies in crowd scenes, i.e., human behaviors anomalous with respect to the rest of the behaviors. In this paper, we introduce a new concept of contextual anomaly into the field of crowd analysis, i.e., the behaviors themselves are normal but they are anomalous in a specific context. Our system follows an unsupervised approach. It automatically discovers important contextual information from the crowd video and detects the blobs corresponding to contextually anomalous behaviors. Our experiments show that the approach works well in detecting contextual anomalies from crowd video with different motion contexts.

Index Terms— Crowd analysis, anomaly detection, clustering

1. INTRODUCTION

In many surveillance systems in public places such as city streets, subway stations, malls, video is recorded depicting the movement of crowds. It would be very useful to locate and recognize hazardous and anomalous human motions from the video to alert system operators. By anomaly detection, we mean to detect motion patterns in the video that do not conform to the expected behavior. However, in an arbitrary video scene it is hard to predefine normal behavior (requiring elaborate work of labeling and training). We can define anomaly as rare or infrequent behavior compared to all other behaviors, which is also referred to as an outlier [1]. This paper will focus on the unsupervised anomaly detection, which aims at automatically mining anomaly behaviors without normal pattern training.

The simplest type of anomaly, which is also the focus of the majority of research [2–6], is to detect an individual behavior instance that is considered as anomalous with respect to the rest of behaviors. This type of anomaly is called point anomaly [1].

However, sometimes the individual behavior itself has similar features with others but it is anomalous in a specific context (e.g., neighborhood); then it is termed as a contextual anomaly [1]. This is not widely studied for crowd analysis in previous research. One recent work [7] proposed an unsupervised framework adopting hierarchical Bayesian models to model activities and interactions in crowded scenes. Anomaly detection could be done for atomic activities (corresponding to point anomalies) and interactions (related to context, but constrained to specific objects).

We also see context information used in other research areas. [8] detected irregularities in images and videos based on an ensemble of spatially related patches. [9] matched objects based on a similar geometric layout of visual patches within a surrounding image region. These patches and their layout are actually a description of the spatial context.

Inspired by these works, we propose an unsupervised approach for detecting contextual anomalies in crowd motion. First, motion features in the crowd scene are represented by spatial temporal patches which are characterized by dynamic texture. All patches are then classified and grouped to blobs that approximately describe position and size of every pedestrian. This is described in Sec. 2. Then based on the spatial layout of pedestrians with different motions, our system automatically discovers important contextual information and detects the blobs corresponding to contextually anomalous behaviors. This is described in Sec. 3. Experimental results are presented in Sec. 4. Finally we conclude the paper in Sec. 5. Our main contribution is introducing the concept of contextual anomaly into the field of crowd analysis, and proposing an approach to automatically detect those contextual anomalies (which could not be achieved by previous works on point anomaly detection).

2. MOTION REPRESENTATION AND CLASSIFICATION

Due to the density of objects in a crowd scene, accurately tracking individual objects is difficult. We characterize the crowd motion by the patch-based local motion representation. Similarly to [6], the non-stationary parts in the video are represented as a collection of spatio temporal patches of dimension $p \times p \times q$, where p (spatial size) and q (temporal size) should be large enough to capture the distinguishing characteristics of the various components of the local motion field. Every patch is characterized by a dynamic texture model [10], which is actually a linear dynamic system defined as

$$
\begin{aligned}
x_{t+1} &= Ax_t + Bv_t \qquad (1)\\
y_t &= Cx_t + w_t, \qquad (2)
\end{aligned}
$$

where $y_t \in R^m$ ($m = p \times p, t = 1, 2, \cdots, q$) is the appearance vector of a patch at each time, which is regarded as an observation drawn from the hidden states $x_t \in R^n$ ($n \ll m$). v_t and w_t are the driving process and the observation noise process, respectively. Given the appearance vectors y_t, this model (x_t, A, C) can be parameterized by a suboptimal (but tractable) approach in [10].

Based on this representation, we can cluster all patches into different motion categories (behaviors). In our work, we adopt the spectral clustering algorithm. The pairwise distance between two patches is defined by Martin distance [11], which is based on the principal angles between the subspaces of the extended observability matrices of the two dynamic textures and can be computed by matrices A and C. One example is shown in Fig. 4. Fig. 4(a) shows one frame of the crowd video, where many people walk in two directions on the road. Fig. 4(b) shows the patch representation and clustering results for this frame. All the non-stationary parts in this video are represented by patches (with dimension $10 \times 10 \times 20$) and all the patches are clustered into two categories (colored in green and blue separately).

3. CONTEXT REPRESENTATION AND ANOMALY DETECTION

At this patch representation level, there is no anomalous motion in the video because every patch falls into one of two normal clusters. However, there exist some contextual anomalies. When we consider the context all through the whole video, we find that in this video most of the time people follow the crowd flow, i.e., they walk in the same direction as their neighbors. Only in very few cases people disobey this rule by walking in the opposite direction of the flow. One example is the green blob at the upper-right corner of Fig. 4(b) surrounded by blue patches, where one pedestrian is walking downwards while pedestrians in his neighborhood all walk upwards. We aim to automatically detect this kind of contextual anomaly.

In this work, we consider the motion context of pedestrians. Ideally, we can segment every pedestrian and find its consisting patches, then use the class label (green or blue) at each pedestrian's neighborhood as the contextual attributes. However, accurate segmentation of people from crowd scenes is a difficult problem, and our goal is not pedestrian boundary detection but motion context representation. Hence, a blob representation that coarsely corresponds to pedestrians, as long as the contextual information is not changed, can serve our purposes.

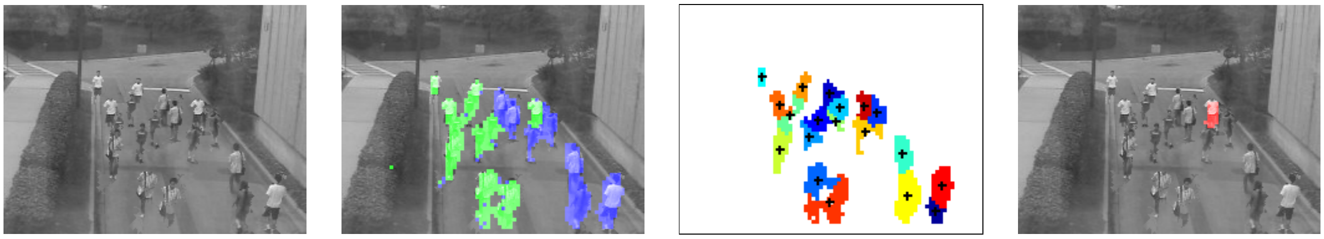
Obviously, the blob representation is related to pedestrian size, therefore it is important to consider the effects of perspective, as in the example video pedestrians closer to the camera appear larger. In this work, we manually find two sizes (in number of patches) of an average person at the nearest and the farthest end, respectively, in the video scene. Then we approximate the size of any pedestrian at any place by linearly interpolating between the two size extremes.

Aided by the estimated pedestrian size, we perform region growing on the patches of different categories. For example, we scan all the green patches from up to down, left to right. A blob is growing from one patch at a certain position in the image to a connected component of patches. It stops growing when the number of patches reaches the estimated size of the pedestrian at this position, or there are no longer any connected patches. In this way, all the green patches in one frame are grouped into green blobs. The similar patch grouping is performed to all blue patches. Fig. 4(c) shows all blobs (both green and blue) in pseudo-color, with the black crosses denoting their centers. Although each blob does not exactly fit the boundary of every pedestrian, it captures the correct category label (green or blue) of each pedestrian and its neighbors. Thus this blob representation has the contextual information of pedestrians and ensures meaningful contextual analysis.

In detail, the motion context for each blob can be defined based on its k nearest neighbors. In the example of Fig. 4, the contextual information of each blob is coded by 1) the category label of itself (0 represents green and 1 represents blue) and 2) the number of neighboring blobs with the same label as itself (from 0 to k). For instance, when we consider the 4 nearest neighbors for each blob, a green blob with 3 green neighbors and 1 blue neighbors is coded as (0, 3). This code represents the blob motion information as well as the motion contextual information.

Once the contextual information is coded for each blob, contextual anomaly can be detected based on its statistics. For a crowd video, we process each frame and gather the contextual codes for all the blobs. The statistics of all the blob codes are represented in a 2-D histogram, where the x and y axis denote the two codes respectively and the z axis is the count of each code. The histogram naturally shows the motion context of the crowd video: high bins are repetitive (thus normal) motion contexts, while low bins are rare (thus anomalous) contexts.

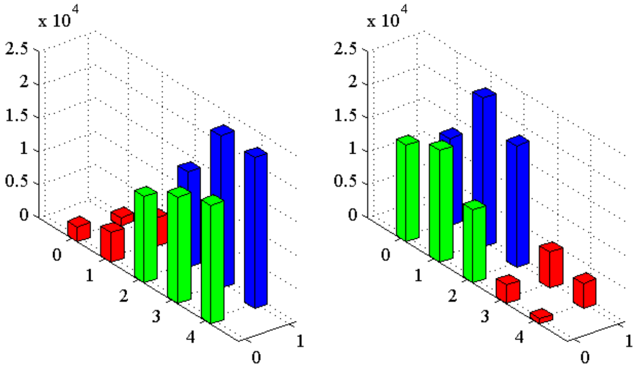
The histogram in Fig. 2(a) shows the statistic of context for the video example in Fig. 4. As most pedestrians in the video walk in the same direction as their neighbors, we have high bins for green (blue) blobs with 2 or more green (blue) neighbors and have low bins for green (blue) blobs with 1 or fewer green (blue) neighbors. By simply imposing a threshold, which can be the average height of all bins, we can dis-



(a) Original video frame (b) Patch classification (c) Blob representation (d) Contextual anomaly

Fig. 1. Example of contextual anomaly detection

cover the low bins that correspond to contextual anomalies (shown as red bins in Fig. 2(a)). In this video, all the blobs that are coded as (0,0), (0,1), (1,0), or (1,1), which correspond to the rare cases of pedestrian walking in the opposite direction of their neighbors, are detected as contextual anomalies and labeled with red in the image as in Fig. 4(d).



(a) For the first video (b) For the second video

Fig. 2. Contextual histogram

4. EXPERIMENTAL RESULTS

In order to show the ability of our method to automatically detect contextual anomaly in different scenarios, we have experimented with two crowd video sequences with different motion context. The first video is composed of 6000 frames, where crowds of people walking as two flows (downwards and upwards) on the two sides of one road. The example shown in Fig. 4 is one frame drawn from this video. More results are shown in Fig. 3, where the top row shows the patch representation of 4 frames selected from the video (two motion categories labeled as green and blue), and the bottom row shows the results of contextual anomaly detection (red blobs). The decision is made based on the contextual histogram shown in Fig. 2(a). The high bins (green and blue bins) denote the normal motion context, i.e., pedestrians (of both flows) walking in the same direction as most of its neighbors. The low bins (red bins) correspond to anomalous motion context, i.e., pedestrians (of both flows) walking in the opposite direction as most of its neighbors.

The second video has similar length and includes many people walking on the same road. However, instead of two clear motion flows shown in the first video, the second video has pedestrians walking in opposite directions intermingled with one another, i.e., random distributed positions. A normal scene is shown in the first image of Fig. 4. This is a totally different scenario than in the first video. As expected, the contextual histogram of the second video (shown in Fig. 2(b)) has high bins corresponding to pedestrians walking in the opposite direction as most of their neighbors, and low bins corresponding to pedestrians walking in the same direction as most of their neighbors. Therefore, the contextual anomaly detected in the second video is no longer those opposite walkers, but those co-walkers. As shown in the results in Fig. 4, a group of people walking upwards together are detected as anomalies. This experiment shows the advantage of our unsupervised anomaly detection approach: it adaptively analyzes contextual information for different crowd videos, and anomaly detection results are always given based on the statistics of the specific crowd scenario, with no a priori knowledge required.

5. CONCLUSION

The main contribution of this paper is to introduce a new concept of contextual anomaly into the field of crowd analysis. Our system focuses on motion context of moving objects and can automatically discover anomalous motions in terms of context (neighborhood motion). This is based on statistical analysis of the contextual information for any given video, thus no priori knowledge is required. Experimental results have been provided for crowd videos with different motion contexts.

6. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, to appear in 2009.

[2] S. Ali and M. Shah, “A lagrangian particle dynamics approach for crowd flow segmentation and stability analy-

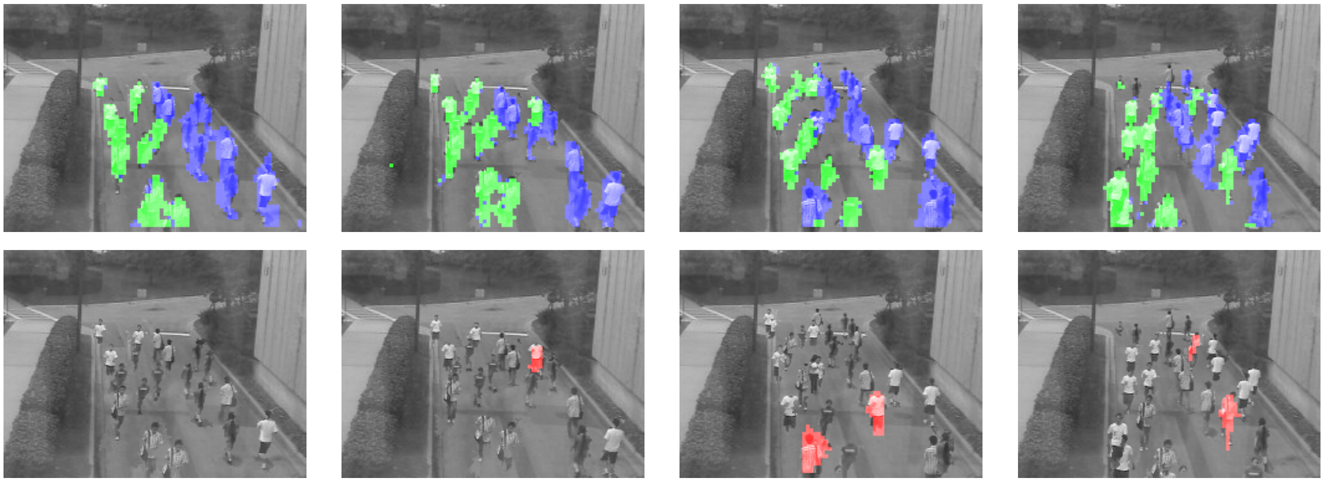


Fig. 3. Anomaly results for the first video

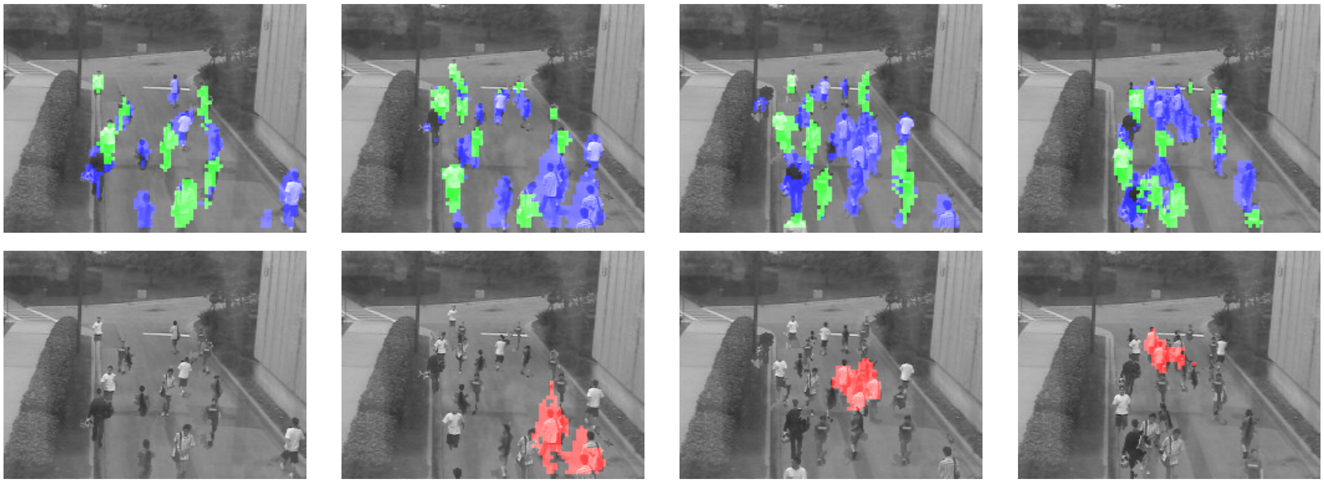


Fig. 4. Anomaly results for the second video

sis," in *Proc. IEEE Conf. on Comput. Vision and Pattern Recognition*, June 2007, pp. 1–6.

[3] M. Hu, S. Ali, and M. Shah, "Learning motion patterns in crowded scenes using motion flow field," in *Proc. IEEE Int'l Conf. on Pattern Recognition*, Dec. 2008, pp. 1–5.

[4] N. Ihaddadene and C. Djeraba, "Real-time crowd motion analysis," in *Proc. IEEE Int'l Conf. on Pattern Recognition*, Dec. 2008, pp. 1–4.

[5] A.M. Cheriyadat and R.J. Radke, "Detecting dominant motions in dense crowds," *IEEE Journal of Selected Topics in Signal Process.*, vol. 2, no. 4, pp. 568–581, Aug. 2008.

[6] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.

[7] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2008.

[8] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *Proc. IEEE Int'l Conf. on Comput. Vision*, Oct. 2005, vol. 1, pp. 462–469.

[9] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. on Comput. Vision and Pattern Recognition*, June 2007, pp. 1–8.

[10] S. Soatto, G. Doretto, and Y. N. Wu, "Dynamic textures," in *Proc. IEEE Int'l Conf. on Comput. Vision*, July 2001, vol. 2, pp. 439–446.

[11] R. J. Martin, "A metric for ARMA processes," *IEEE Trans. Signal Process.*, vol. 48, no. 4, pp. 1164–1170, Apr. 2000.