Cumhuriyet Üniversitesi Fen Fakültesi
Fen Bilimleri Dergisi (CFD), Cilt:36, No: 3 Özel Sayı (2015)
ISSN: 1300-1949

Cumhuriyet University Faculty of Science
Science Journal (CSJ), Vol. 36, No: 3 Special Issue (2015)
ISSN: 1300-1949

A novel method for localizing abnormal behavior in crowded scene

Hamidreza RABIEE[1], Javad HADDADNIA[2*], Omid RAHMANI SERYASAT[1]

[1]*Department of Electrical and Computer Engineering, Hakim Sabzevari University, Sabzevar, Iran*

[2]*Associate Professor, Electrical and Computer Engineering Department, Hakim Sabzevari University, Sabzevar, Iran*

Received: 01.02.2015; Accepted: 06.06.2015

Abstract. Computer vision algorithms have played a vital role in video surveillance systems to detect surveillance events for public safety and security. Even so, a common demerit among these systems is their unfitness to handle divers crowded scenes. In this paper, we have developed algorithms which accommodate some of the challenges encountered in videos of crowded environments to a certain degree. Unlike many approaches that use optical flow, that estimates motion vectors only from two successive frames, we made our descriptor over long-range motion trajectories which is named short trajectorys in the paper. This paper presents a novel video descriptor, referred to as Histogram of Short trajectorys, for detecting abnormal conditions in crowded scenes. Specifically, we extract 2-d histogram from magnitude and orientation matrixes which describe the motion patterns expected in each cubid. We classify frames as normal and abnormal by using machine learning methods.

Keywords: abnormal behavior, crowded scene

1. INTRODUCTION

Safety and security have always been a major issue for shopping centers, banks, official buildings, enterprises, etc. Nowadays, nearly everybody looks for a way to keep its belongings safe and secure.

Improvements in new technologies are making possible the development of many systems of safeguarding and surveillance for all necessities and budgets. Video surveillance is commonly used in security systems, but requires more intelligent and robust technical approaches. Such systems, which used in airports, subways, banks, concerts, cinema halls, sporting events, schools, supermarkets, parking places, hospitals, hotels, town centers or other private/public spaces, can bring security to a high level. The scientific challenge is to invent and implement automatic systems for obtaining detailed information about the activities and behaviors of people or vehicles observed by sensors (e.g., cameras). Automatic video surveillance is attractive because it promises to replace more costly option of staffing video surveillance monitors with human observers.

Figure 1 demonstrates the kind of scenarios we consider here by preparing some instances from our database of both violent and non-violent crowd behavior.

* Corresponding author. *Email address: jhaddadnia@yahoo.com*

Special Issue: The Second National Conference on Applied Research in Science and Technology

Figure 1. Examples of violent (top 2 rows) and non-violent (bottom row) crowd behavior in "real-world" videos.

1.1. Related works

Interest-point based methods begin by first detecting space-time key-points [6,16]. Descriptive information is then extracted at each of these points using one of several space-time descriptors (see for example: [8, 12, 13]). A video can then be represented using, e.g., Bag-of-Feature techniques (as in [13, 14]). These methods are often very resilient to camera motion and have been shown to provide excellent performance on a number of challenging benchmarks [11, 13, 14], however, when videos contain too few space-time interest points (e.g., little motion) or too much motion (as in our scenarios), they may fail to efficiently provide meaningful representations.

The alternative of considering whole frames, or frame parts, often builds on dense flow estimation between successive frames [3, 4] or high-level appearance models [1].

Related to crowd videos are the methods of [9] and more recently Rodriguez et al. [18]. Both these methods are data-driven and require matching parts of the query video frame segments in [9] and spatiol-temporal cubes in [18] to exemplars in a pre-collected database. Searching the database for matching exemplars would be impractical for the applications considered here.

Unlike the mentioned works, our approach is based on short trajectories, which are typically short temporal sequences of tracked points, usually extracted using the KLT method [19]. The points to track can be different points [22, 23], randomly selected points or densely distributed points on a grid [21]. Short trajectory-based methods can be viewed as a trade-off between object- and holistic-based methods. The scene is modelled using trajectories of a set of points without considering to their fist place (i.e. generated by person A or B), which is a much simpler work. From short trajectories a new spatio-temporal feature is built: for the analyzed sequence and for the found short trajectories, we estimate the Histogram of Short trajectory, which shows the huge motions over a short period of time within a region of the scene. specifically, it describes the flow patterns within a sub-spatiotemporal place.

assuming that abnormalities are outliers, we applied two standard methods for classification, i) when just normal data is in hand for training we used Latent Dirichlet Allocation (LDA) [5], ii) if abnormal training data is available too, we employed Support Vector Machines.

2. HISTOGRAM OF SHORT TRAJECTORY

Short trajectories [18] are closely spatio-temporal elements of moving firm objects. They show fragments of a whole trajectory of individual point displacement, produced by frame-wise connection between point localization results in the near frames. Here, we examine the effectiveness of short trajectories using histogram of them for detect abnormal behaviour in crowded scenes. In details, a short trajectory can be demonstrated as a succession of points in the video space:

$$tr^{(k)} = (q_1^k, ..., q_t^k, ..., q_T^k) \quad (1)$$

In above equation q_t^k are two dimension points (x_t^i, y_t^i) of a sequence over a frame.

Short trajectorys are reached through choosing regions of interest in frames by a feature detector and track them in time until occlusion or clutter led them to be lost. On this basis, this approach can be views as a trade-off between object based methods and holistic methods. Instances of short trajectorys are shown in figure 2.

Figure 2. Examples of Short trajectorys extracted from sample dateset (left) and from UCSD dataset (right).

In a crowded scene there could be different patterns of motions in diffrenet regions, so a descriptor must be designed which can encompasses statistics of motion of short trajectories passing through spatio-temporal 3D cells. Fig. 3(a) shows a schematic view of our method.

At first step using OpenCV code we extracted all short trajectories in a given video stream. Following that, we used HoG algorithm to detect the salient points and then tracked them using KLT algorithm. [20]. So, for every frame, new short trajectories are started at the location of the detected interest points. we have N short trajectories for a given video stream, $\zeta = \{tr^1, ..., tr^n, ..., tr^N\}$. The length of a short trajectory is depended on the intensity of motion-patterns in the scene and the relative position of the camera. This leads to large number of short trajectories that can completely describe the motion patterns of the observed scene.

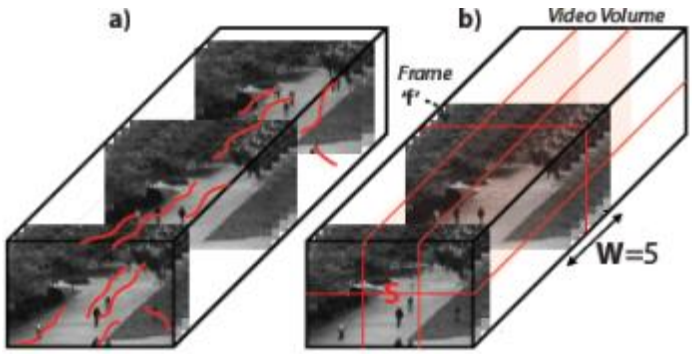

Figure 3. a) Short trajectories b) To compute HST the video volume is divided in non-overlapping spatial parts. Then for every frame we considered a temporal window stride of W frames.

In the second step, as showed in Fig.3(b), spatio-temporal 3D cells of specific size are extracted with splitting the videos. These are Overlapped only in temporal domain. For each 3D cell we calculate orientation and magnitude of all short trajectories that passing through it as following:

$$\beta^{i,s} = \tan^{-1} \frac{y_{end}^{i,s} - y_{start}^{i,s}}{x_{end}^{i,s} - x_{start}^{i,s}} \quad (2)$$

$$Mag^{i,s} = \max_{T,T+1 \in v} \{\sqrt{(x_{T+1}^{i,s} - x_{T}^{i,s})^2 + (y_{T+1}^{i,s} - y_{T}^{i,s})^2}\} \quad (3)$$

Apex (i,s) is referred to the portion of the short trajectory i that pass through 3D cell s and we show $x_{start}^{i,s} - y_{start}^{i,s}$ and $x_{end}^{i,s} - y_{end}^{i,s}$ as entry and exit points of short trajectory i in/from 3D cell s .

Fig.4 demonstrates this concept for a short trajectory.

Finally, M and O values are extracted using magnitude and orientation quantization. With counting how many times a specific orientation-magnitude pair is observed the bins of a histogram $H_{Mag,\beta}$ are populated. The resulting 2D histogram is called histogram of short trajectory (HST).

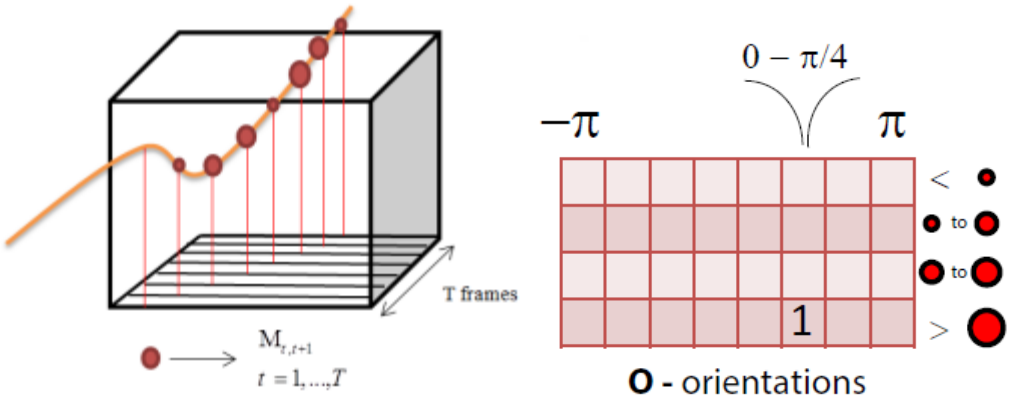

Figure 4. The process of HST formation. The magnitude of the short trajectory, here illustrated by red circles, is calculated in every point (frame) of the temporal window.

A descriptor at frame level $H_{\beta,Mag}^{s,f}$ can be computed by applying a sliding window to frames. Histogram at frame f is based upon 3D cells that ranges from $f - \frac{Win}{2}$ to $f + \frac{Win}{2}$. The major difficult is that abnormalities are detected with some delay because of i) the need for future frames and ii) the frame descriptor still encompasses data from pas frames when abnormality begins. But, the time latency is at most $0.3 - 0.6$ seconds and is negligible. Fig5. Demonstrates the projection of the image of HST descriptors for 3 frames fro, USCD dataset.

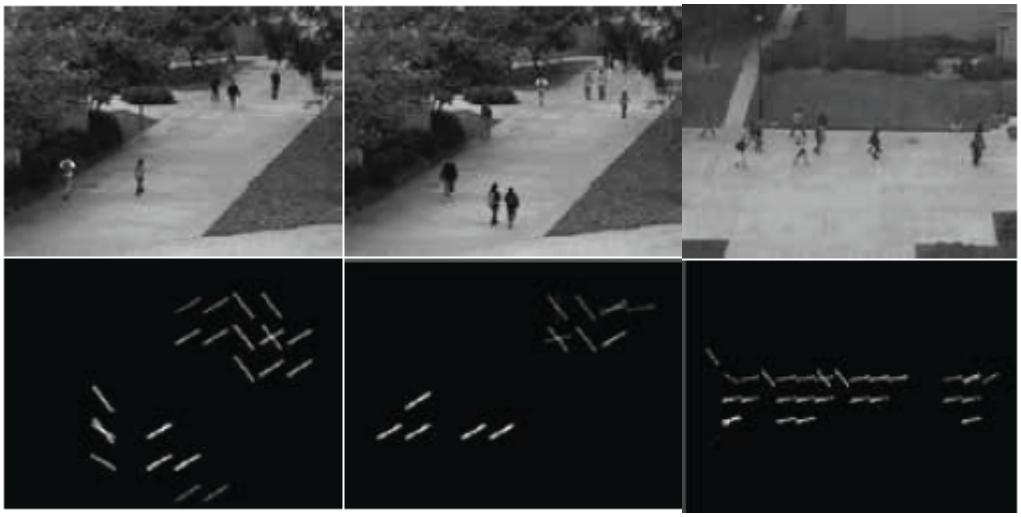

Figure 5. (top) 3 frames for ped, (bottom) the HST descriptor averaged over the orientations and projected in every part on the image plane, shows the expected motion in each sub-window.

3. ABNORMALITY DETECTION USING LDA

In crowded scene we cannot suppose that abnormal footage is available at the training time because it would be very hard to prepare huge training set all probable abnormal behaviour. The normal situations on the contrary are simply in hand. In this case, a better choice to model abnormal behaviour is using generative models to mathematically learn "what the conception is" and labelled any video which is not classified as normal to abnormal.

The LDA will define what is normal between features which occur simultaneously. With having 2D histogram $H^{s,f}_{\beta,Mag}$ for each frame we built the LDA training corpus regarding below detection scheme:

3.1. Each-frame, Each-independent sector

Here we learn a LDA model each-sector which is independent. The training collection of per sector is constructed by

$$C^s = \{H^{s,f}_{\beta,Mag}\}^F_f \tag{4}$$

Sub-region level abnormality detection using this method is straightforward. Expanding the collection is possible for instance in non-uniform portions and spatio temporal pyramids.

Once C in collected, we compute the histograms and let the LDA to learn the set of topics that describe what normality is. For iteratively maximizing a bound on the information log-likelihood we applied Expectation-Maximization algorithm (EM).

$$L(C|\alpha,\gamma) = \sum_f \log p(C^f|\alpha,\gamma) \tag{5}$$

In above equation α is the Dirichlet prior over topic combination, and γ characterizes the motion patterns related to per topic.

We can estimate the log-likelihood of an unseen test frame $L(C^{f}_{unobserved} \left| \hat{\alpha}, \hat{\gamma} \right.)$ and label the frames as normal or abnormal based upon a fixed threshold on the evaluated likelihood.

3.2. Crowd datasets

Our method is tested by three standard dataset: UMN[15],Violence-In-Crowds[7], and USCD[21].

The third dataset includes videos of a crowded pedestrian walkway with manually frame-level ground truth. Because of having two scenes the dataset is divided into two subsets. The first contains clips of 150*234 pixels, that shows group of people walking toward or away from the camera this dataset is called “ped1”. The second dataset has resolution of 230*340 pixels and demonstrates horizontally moving pedestrians and called “ped2”. (see Fig. 6).

The violence-in-crowd dataset is extracted in real-world, video footage of crowd violence with protocols designed to test violent and non-violent classification. We can test standard SVM and HST descriptor with it.

We also used some sample of UMN dataset including eleven different scenarios of a panic and normal conditions in two different outdoor and indoor scenes.

3.3. HST Parameter Tuning

Like other descriptors some of constants and parameters have to be tuned. For instance tessellation of the frame S, the quantization bins O and M, the temporal window W and LDA’s topics.

We quantized short trajectories orientation in 8 uniform bins and temporal window size was 11. For magnitude of short trajectories quantization levels are {3, 5, 7, 12, 15, 30}.for magnitudes quantization intervals we used linear spacing from o up to maximum value of magnitude found in training set. Three different spatial tessellations were used in our work, we called corase S=2*3, medium S=4*6 and fine S=8*12. The LDA topics are considered 2,4,6,…,80 and learning was repeated 5 times.

4. EXPERIMENTAL RESULTS

In this section we compare HST with some methods and descriptors particularly with the approach proposed in [21] that models motion flows in crowds by the use of Social Force Model and MDT [15], and the leading methods based upon optical flows [2,10,7].

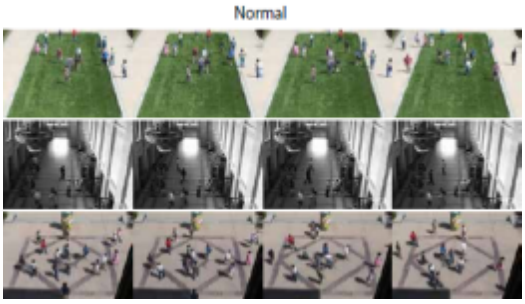

Figure 6. Normal and abnormal frames from the three scenarios of the UMN dataset(Top). Here abnormality is fear with people running away. Normal and Violent crowd from the Violence-in-crowds dataset. Videos are from various scenes(Left).

5. DETECTION PERFORMANCE

Detection of abnormal behaviours: we regarded UCSD dataset and its standard train-test partition for this experiment. We used the likelihood of test frames to evaluate the Area Expected Error Rate and under the Curve measures.

Each-frame, Each-independent sector Approach is fruitful to regard simultaneous movements in various areas of the scene. Some poor performance occurred because of perspective distortion and over-train LDA in some regions. This problem can be solved using adjustment LDA or with several rough estimations of the scene geometry, but it is not in the scope of our work.

In table.1 our results are compared with other methods. Although we cannot highlight a specific winner across all the model configurations we focused ourselves to acknowledge how the analysis of short trajectories proposed here (HST) produces performance and certainly outperforms schemes based upon interactions like SFM [15] and optical flow [2]. HST resulted in ac AUC of over 0.99, likewise some other methods but work better than social force model [15] (AUC 0.96) and optical flow (AUC 0.92).

Table 1. Expected Error Rates on USCD dataset by use of standard testing protocol. We report here the best result obtained by all the works (as previously done by [21]).

method	EER
Social force model [13]	36.6%
Mixt. Dyn. Textures [22,11]	23%
HST	**20.5%**

6. CONCLUSION

In this paper we proposed Histogram of short trajectory (HST) as a new video descriptor for detecting behaviour abnormality in crowded scene. This new descriptor gets magnitude and orientation of a single feature. Results show that how the analysis of short trajectories can be effective in abnormality detection. This is effective as we just used standard classifiers and didn't mix descriptors. We also tried to evaluate HST with dense trajectories [23] but results were not good in only static camera scenarios.

REFRENCES

[1] M. Abdelkader, W. Abd-Almageed, A. Srivastava, and R. Chellappa. Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. CVIU, 115(3):439–455, 2011.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using mulitple fixedlocation monitors. IEEE Trans. Pattern Anal. Mach. Intell, 30(3), 2008.

[3] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. TPAMI, 32(2):288–303, 2010.
[4] R. Basri, T. Hassner, and L. Zelnik-Manor. Approximate nearest subspace search. TPAMI, 33(2):266–278, 2011.
[5] D.Blei, A.Ng, and M.Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.
[6] P. Doll´ar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In VS-PETS, pages 65–72, 2005.
[7] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In CVPR Workshops, pages 1–6. IEEE, 2012.
[8] M. Kaˆaniche and F. Br´emond. Recognizing gestures by learning local motion signatures of HOG descriptors. TPAMI, 2012.
[9] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In ICCV, pages 1–8, 2007.
[10] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates, 2009. CVPR.
[11] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. TPAMI, (99):1–1, 2012.
[12] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In CVPR, pages 2046–2053, 2010.
[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In CVPR, pages 1–8, 2008.
[14] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos 'in the wild'. In CVPR, pages 1996–2003, 2009.
[15] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model, 2009. CVPR.
[16] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In CVPR, pages 1–8, 2007.
[17] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In ECCV, pages 577–590, Berlin, Heidelberg, 2010.
[18] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In ICCV, pages 1389–1396, 2009.
[19] J. Shi and C. Tomasi. Good features to track, 1994. CVPR.
[20] C. Tomasi and T. Kanade. Detection and tracking of point features, 1991. Intl Journal of Computer Vision.
[21] H. Wang, A. Kl¨aser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In IEEE Conference on Computer Vision & Pattern Recognition, pages 3169–3176, Colorado Springs, United States, June 2011.
[22] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets, 2011. CVPR.
[23] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors. learning a mixture model of dynamic pedestrian-agents, 2012. CVPR.