Pedestrian Anomaly Detection and Localization

Faeghe Negini, K.N.Toosi University of Technology, Iran
Ali Ahmadi, K.N.Toosi University of Technology, Iran
Pejman Hashemi Bakhtiar, K.N.Toosi University of Technology, Iran

Received: August 26, 2015
Accepted: September 28, 2015
Published: October 30, 2015

Abstract: A run-time improvement for anomaly detection in crowded scenes is proposed. This framework is a combination of temporal anomaly detection and edge detection methods. The presented detector is based on a video representation, it uses a set of models for normal crowd behavior containing mixtures of dynamic textures which consists of both appearance and dynamics. After detecting temporal anomaly, edge detection is performed on processed frames, and temporal anomaly detection output contours will be fit to detected edges. This proposed method allows better accuracy in localization without significant additional workload.

Keywords: Video surveillance, Anomaly detection, Crowded scene, Dynamic texture

1 Introduction

The recent population boom has made crowd phenomenon more frequent and has created new needs for crowd analysis. Specifically, the behavioral analysis of crowded scenes is of great interest with large number of applications, such as (Junior 2010):

1) Crowd management

Crowd analysis can be used for amplifying crowd management strategies, to avoid crowd related disasters and ensure public safety.

2) Public space design

It can provide guidelines for the design of public spaces.

3) Virtual environments

It can be used to validate or increase the performance of the mathematical models used in crowd simulations.

4) Visual surveillance

It can be used for automatic detection of anomalies and raise alarms. Moreover, the ability to track objects in a crowd could help the security guard to catch suspects.

5) Intelligent environments

In some intelligent environments which involve large groups of people, crowd analysis can be used to take a decision for assisting the crowd or an individual in the crowd.

The study of human behavior is a subject of great scientific interest, especially for intelligent visual surveillance it has gotten more research attention and funding following increased global security concerns and increasing need for effective monitoring of public places such as airports, railway stations, shopping malls, crowded sports arena, military installations, etc.

To recognize human activity is to automatically analyze ongoing activities from an unknown video.

Usually, detecting, recognizing, or learning favorite events which is defined as ‘abnormal behavior’ or ‘abnormality’ is one of the major goals. The term “behavior” is generic and refers to the noticeable actions of agents such as persons, or other moving objects in the scene. Such salient manners are because they are different from the normal patterns in that context. Thus anomalies are temporal or spatial outliers events not conforming to learned patterns. They “stand out” as different relative to the context of their surrounding in space and time (Zhan 2008).

We can view abnormal behavior detection as a type of high level operation of image understanding, where logical information is extracted from input image sequences and used to model behavior. This figure shows sketch of the general process (Popoola 2012).

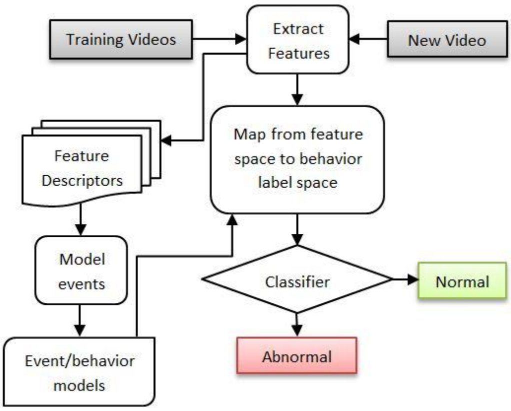

Figure 1. General process of feature based modeling and detection of anomalies in video sequences. *(Popoola 2012)*

The research in abnormal behavior detection based on both prior knowledge are used and human affecting on the learning process are categorized as supervised, unsupervised and semi-supervised.

Whereas in this paper unsupervised method is used, a review on existing approaches in this area is given at the following (Popoola 2012).

Several approaches have been proposed to resolve the problem of abnormality detection that are categorized based on types of scenes and learning models that they use. Most of the proposed approaches for tracking objects in crowded scenes however, requires effective background subtraction and is limited by factors such as occlusion and shadows, so limiting its advantage to encode compound behavior in real world event of anomalies (Basharat 2008, Siebel 2002, Zhang 2009l). Moreover, it is sensitive to tracking errors even though they occur in a few frames. This approach also fails when modeling crowded and complicated scenes (Mahadevan 2010).

Various authors have proposed alternative motion representations that avoid tracking. The most popular is dense optical flow, or some other form of spatio-temporal gradients (Adam 2008, Kim 2009, Mehran 2009).

All of these approaches concentrate exclusively on movement information, due to appearance variety of objects appearance information is often ignored which make them unbreakable against abnormalities without motion outliers, but these approaches have problems in crowded scenes where background is dynamic, the scene is cluttered or has complicated occlusions. Some advanced representations have worked on both appearance and movement. Boiman and Irani (2007) have used spatio-temporal patches and declare regions that cannot be reconstructed using data from previous frames as abnormal. Spatio-temporal gradients gave been proposed in (Kratz 2009), where their statistics are modeled with a coupled HMM to detect abnormalities in densely crowded scenes (Mahadevan 2010).

In this work spatio-temporal patches are used that have extracted from video cells, then each cell is modeled by one mixture of dynamic textures(MDT). This space-time local features have been particularly popular because of their reliability under noise, camera jitter, illumination changes, and background movements (Aggarwal 2011).

In this paper, anomalies are considered based on events of low-probability in relation to a model of normal crowd behavior. Then these results were combined with outcome of edge detection process to improve system accuracy.

The evaluation is based on a dataset of crowded scene which contains video sequences of a college campus walkway with crowds with naturally varying densities. It contains abnormal events that occur naturally, e.g. bicycle riders, cars, a person who walks through the grass. In the end the proposed approach is compared with previous methods.

2 Anomaly Detection

2.1 Mixtures of dynamic textures

According to Chan (2005), a mixture of dynamic textures model is a collection of videos consisting of different visual processes as samples from a set of dynamic textures.

2.1.1 Dynamic texture

A dynamic texture is a generative video model defined by the following linear dynamical system (LDS) equations. It consists of a random process containing an observed variable y_t, which encodes the appearance of video frame at a specific time, and a hidden state variable x_t, which encodes the dynamics of video over time Chan (2005).

$$\begin{aligned} x_{t+1} &= Ax_t + v_t \\ y_t &= Cx_t + w_t \end{aligned} \tag{1}$$

where $x_t \in R^n$ and $y_t \in R^m$, $(n \ll m)$. The parameter A $\in R^{n\times n}$ and $C \in R^{m\times n}$ are the state transition and observation matrices respectively. The driving noise process is $v_t \sim N(0, Q)$ with $Q \in R^{n\times n}$, and observed noise process is $w_t \sim N(0, R)$, with $R \in R^{m\times m}$, where $N(\mu, \Sigma)$ is a Gaussian distribution with mean μ and covariance Σ. The initial condition is given by $x_1 \sim N(\mu, S)$. The dynamic texture is completely specified with the parameters $\Theta = \{A, Q, C, R, \mu, S\}$

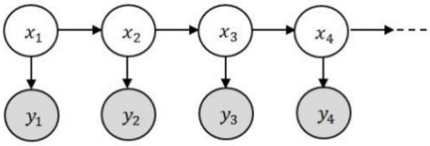

Figure 2. Dynamic texture *(Chan 2005)*

2.1.2 Mixture of dynamic textures

The video sequence y_1^τ is produced from one of k dynamic textures and with nonzero prior probability $\alpha_j$of occurrence, then component probabilities are $\{\alpha_1, \dots, \alpha_k\}$ with $\sum_{j=1}^{k} \alpha_j = 1$and dynamic texture components of parameters $\{\Theta_1, \dots, \Theta_k\}$, a video sequence is taken by: producing a mixture component index z and observation y_1^τ form the dynamic texture component of Θ_z and an observation y_1^τ from the dynamic texture component of parameters Θ_z (Chan 2005).

The probability of a sequence $\boldsymbol{y_1^\tau}$ under this model is

$$p(y_1^\tau) = \sum_{j=1}^{k} \alpha_j p(y_1^\tau | z = j) \tag{2}$$

where $p(y_1^\tau | z = j)$ is the class conditional distribution of the jth dynamic texture, that is, the dynamic texture component parameterized by $\Theta_j = \{A_j, Q_j, C_j, R_j, \mu_j, S_j\}$. The generative model for the mixture of dynamic textures is

$$x_{t+1} = A_z x_t + v_t$$

$$y_t = C_z x_t + w_t \tag{3}$$

where the random variable $z \sim multinomial(\alpha_1, \dots, \alpha_k)$ indexes mixture components from which the observations are taken, the initial condition is given by $x_1 \sim N(\mu_z, S_z)$ and the noises are $v_t \sim N(0, Q_z)$ and $w_t \sim N(0, R_z)$. The conditional state distribution and the conditional state observation, given the component index z are

$$\begin{aligned} p(x_1|z) &= G(x_1, \mu_z, S_z) \\ p(x_t|x_{t-1}, z) &= G(x_t, A_z x_{t-1}, Q_z) \\ p(y_t|x_t, z) &= G(y_t, C_z x_t, R_z) \end{aligned} \tag{4}$$

and the joint distribution is

$$p(x_1^\tau, y_1^\tau, z) = p(z)p(x_1|z) \prod_{t=2}^{\tau} P(x_t|x_{t-1}, z) \prod_{t=1}^{\tau} P(y_t|x_t, z) \tag{5}$$

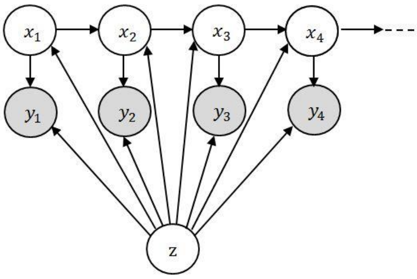

Figure 3. Mixture of dynamic texture. *(Chan 2005)*

2.1.3 Parameter estimation using EM

The EM algorithm is a method for learning of Dynamic Textures and estimation the parameters of a probability distribution when the distribution dependents on hidden variables like missing data in this case. given a set of independent and identically distributed(i.i.d) video sequences $D_i = \{y^{(i)}\}_{i=1}^{N}$, Maximum likelihood estimates (MLE) of the parameters of an MDT $p(y; \Theta)$ of K components, that are learned with the EM algorithm. the hidden variables in this model consists of 1)the assignment $z^{(i)}$ of each sequence to a mixture component and 2) the hidden state sequence $x^{(i)}$ that generates $y^{(i)}$ (Chan 2005).

$$\Theta^* = arg \max_{\Theta} p(D_i; \Theta) = arg \max_{\Theta} \sum_{i=1}^{N} log\, p(y^{(i)}; \Theta) \tag{6}$$

The EM solution is an iterative procedure that alternates between estimating the missing information with the current parameters and computing new parameters given the estimate of the missing information. The EM iteration is between two steps

E-Step: $\quad Q(\Theta; \hat{\Theta}) = E_{D_h|D_i;\hat{\Theta}}[log\, p(D_c; \Theta)]$

M-Step: $\quad \hat{\Theta}^* = arg \max_{\Theta} Q(\Theta; \hat{\Theta}) \qquad (7)$

where the hidden data D_h consists of the hidden variables $\{x^{(i)}\}_{i=1}^{N}$and $\{z^{(i)}\}_{i=1}^{N}$, and the complete data $D_c = D_i \cup D_h$. The assignment variable $z^{(i)}$ is represented by a vector $z_i \in \{0,1\}^K$, such that $z_{i,j} = 1$ if and only if $z^{(i)} = j$.

Each dynamic texture component Θ_j was initialized by using the suboptimal learning method on a random video sequence from the training set. The component probabilities were initialized to a uniform distribution, $\alpha_j = \frac{1}{k}$. Since the EM algorithm can terminate on a local minimum, the algorithm was run several times using different initialization seeds, and parameters which best fit the training data (in the maximum likelihood sense) were kept. Finally, the covariance matrices Q, S and R were regularized by forcing their eigenvalues to be larger than a minimum value, and by restricting S and R to be diagonal (Chan 2005).

2.2 Temporal Anomaly Detection

Abnormality detection operates based on the background subtraction method of GSG(Generalized Stauffer-Grimson) (Stauffer 1999). This method relies on a Gaussian mixture(GMM) at each image location. For abnormality detection the GMM is replaced by MDT, and the pixel-wise grid is replaced by one with a displacement of a defined size. Each grid location defines the center of a video cell. Spatio-temporal patches are extracted from each cell, and a MDT is learned through training phase. The cell sizes are not acutely important (Mahadevan 2010).

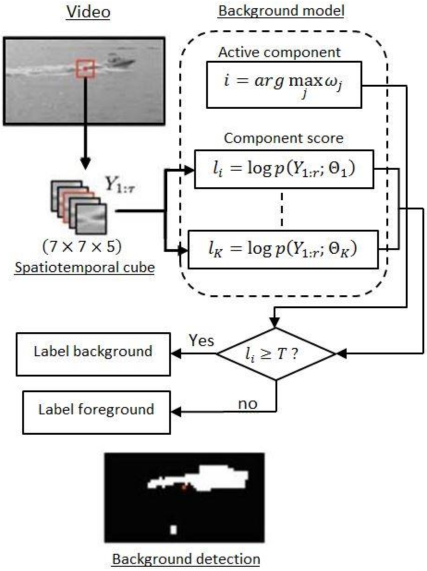

Figure 4. Overview of generalized Stauffer-Grimson background modeling dynamic textures *(Chan 2011).*

In next phase, given a patch y_1^τ, the hidden state sequence, X_1^τ under this MDT model is estimated, and its log-likelihood under the mixture model $p_{X|Y}(X_1^\tau | y_1^\tau)$ is computed with a Kalman smoothing filter. Patches of low probability under the cell MDT are considered abnormalities. The temporal abnormality map at location l is the negative log-likelihood of the state sequence estimated from the patch centered at l (Mahadevan 2010).

$$A_{temporal}(l) = -log\big(p_{X|Y}(X_1^\tau|y_1^\tau);\theta_l\big) \tag{8}$$

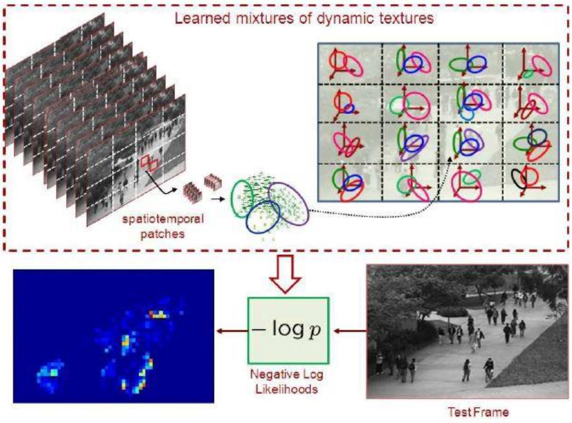

Figure 5. Learning MDTs for temporal abnormality detection. For each region of scene, an MDT is learned during training. At test time, the negative log-likelihood of the spatio-temporal patch centered at location l is computed using the MDT whose region center is closest to l. *(Mahadevan 2010)*

2.3 Edge Detection

This method is an unsupervised edge detection based on the computational edge detection approach introduced by Canny. It is a simple and computationally cheap technique that achieves non-trivial results. This technique uses the highly efficient ADM algorithm to generate the initial edge image, and uses a subsequent modified non-maximal suppression scheme to optimize the edge output resulting in the final edge map without operator intervention of any kind (Ray 2013).

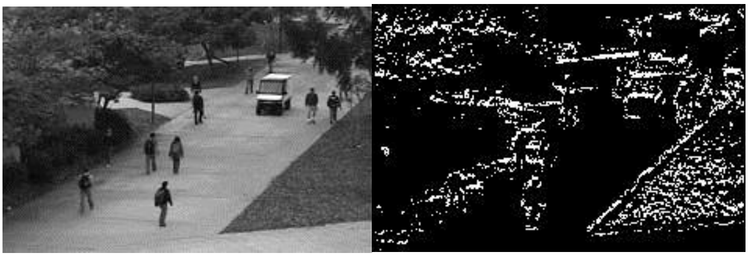

Figure 6. Edge Map of one frame with optimized threshold

2.4 Proposed Method

In this proposed method, anomaly detection is accomplished in three steps:

1- First, temporal anomaly detection procedure is performed as previously described, to identify existing anomalies in each frame.

2- within the boundries of the contour of the region containing an anomaly event determined by the previous step, edge detection is performed to find a more accurate boundry for the subject with anomaly.

3- the contour of the region containing an anomaly event determined in step one is reduced to the boundries identified by the detected edges, to increase the accuracy for our anomaly detection region.

Therefore we can more accurately determine the detected anomaly boundries.

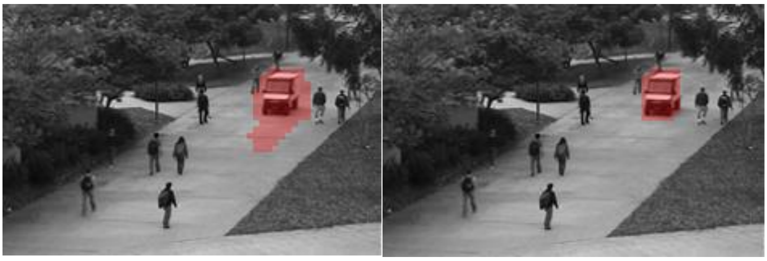

Figure 7. Sample of anomaly detection using MDT approach (left) and proposed approach (right)

3 Evaluation Procedure

In anomaly detection one 'Pixel Level' criteria is used to measure the accuracy of the method. It is based on true-positive rates (TPR) and false positive rates (FPR). A frame consists of anomaly is a positive, otherwise a negative (Mahadevan 2010):

3.1 Pixel level anomaly localization

Each frame recognized by this method as abnormal is compared in pixel level with groundtruth. If at least 40% of the truly anomalous pixels are detected, the frame is assumed as detected correctly and is a true positive, otherwise is a false positive (Mahadevan 2010).

Two measures combined to determine a Receiver Operating Characteristic (ROC) curve of TPR versus FPR.

4 Anomaly Detection Performance

Enhancing localization has been one of our goals in this study.

Following table compare the performance of the tested abnormality detection algorithms. The equal error rate of detection for anomaly localization task is shown in Table 1.

Table 1. Anomaly Localization Experiment: Rate of Detection.

	MPPCA	*MDT(Tmp)*	*MDT(Tmp+Edg) Proposed*
Localization	*18%*	*30%*	*45%*

This is a Real-Time system that needs first training for one time and then works for each new frame immediately. After training phase the mixtures of dynamic textures for videos of frame size 160×240 , the testing time per frame is about 8secs on a standard Pentium machine with 3GHz CPU and 8GB RAM that is improved from earlier proposed systems.

5 Conclusion

It is observed this proposed complex method outperforms similar methods in important aspect of localization accuracy.

the ROC curves for anomaly localization are shown in Figure 8, and the Rate of Detection value is tabulated in Table 1. Example of one frame with anomalies detected by proposed approach are shown in Table 1.

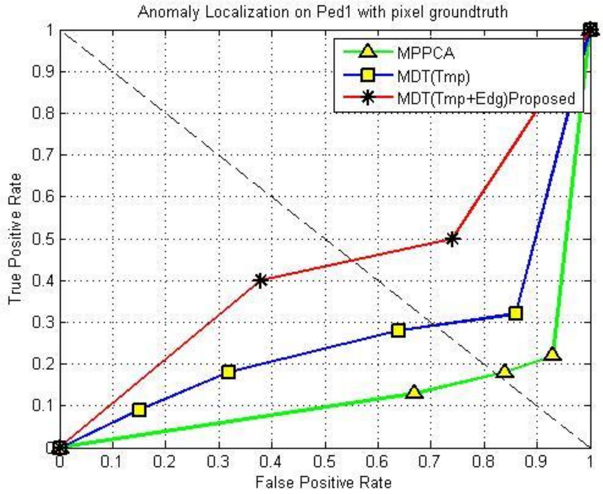

Figure 8. Approaches performance for anomaly localization in pixel level

REFERENCES

Adam, A., Rivlin, E., Shimshoni, I. and Reinitz, D. 2008. *Robust real-time unusual event detection using multiple fixed-location monitors*. PAMI, 30(3):555–560, March.

Aggarwal, J.K. and Ryoo, M.S. 2011. *Human activity analysis: A review,* Journal ACM Computing Surveys (CSUR) Volume 43 Issue 3, Article No. 16, April.

Basharat, A., Gritai, A. and Shah, M. 2008. *Learning object motion patterns for anomaly detection and improved object detection*. In CVPR, pages 1–8.

Boiman, O. and Irani, M. 2007. *Detecting irregularities in images and in video*. IJCV, 74(1):17–31, August.

Chan, A.B., Mahedevan, V., and Vasconcelos, N. 2011. *Generalized Stauffer-Grimson Background Subtraction for Dynamic Scenes*, Machine Vision and Applications, Volume 22, Issue 5, pp 751-766, September.

Chan, A.B., and Vasconcelos, N. 2005. *Mixtures of Dynamic Textures*, Proc. IEEE Int’l Conf. Computer Vision, vol. 1, pp. 641-647.

Junior Silveira Jacques, J. C., Musse, S. R. and Jung, C.R. 2010. *Crowd Analysis Using Computer Vision Techniques*, on IEEE Signal Processing Magazine, vol. 27,no. 5, pp. 66-77.

Kim, J. and Grauman. K. 2009. *Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates*. In CVPR, pages 2921–2928.

Kratz L., and Nishino, K. 2009. *Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models*. In CVPR09, pages 1446–1453.

Mahadevan, V., Li, W., Bhalodia, V. and Vasconcelos, N. 2010. *Anomaly Detection in Crowded Scenes*, IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco.

Mehran, R., Oyama, A. and Shah, M. 2009. *Abnormal crowd behavior detection using social force model*. In CVPR, pages 935–942.

Popoola, O.P. and Wang, K. 2012. *Video-based abnormal human behavior recognition—A review*, on Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, Volume:42 , Issue: 6, Nov.

Siebel, N. and Maybank, S. 2002. *Fusion of multiple tracking algorithms for robust people tracking*. In ECCV, page IV: 373 ff.

Stauffer, C. and Grimson, W. 1999. *Adaptive background mixture models for real-time tracking*. In CVPR, pp. 246–52.

Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A. and Xu, L.-Q. 2008. *Crowd analysis: A survey,* Machine Vis. Applicat., vol. 19, no. 2, pp. 345–357.

Zhang, T., Lu, H. and Li S. 2009. *Learning semantic scene models by object classification and trajectory clustering.* In CVPR, pages 1940–1947.

Ray, K. 2013. *Unsupervised edge detection and noise detection from a single image*, Pattern Recognition, Volume 46, Issue 8, August, Pages 2067–2077