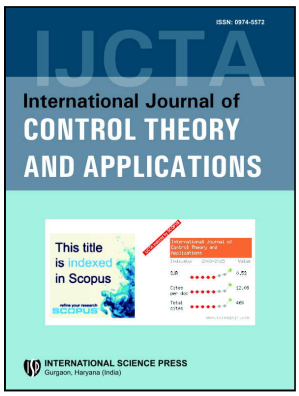

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 13 • 2017

Intelligent Vision Based Surveillance Framework for ATM Premises

Monika Pandey[1], Vishal Sanserwal[1] and Vikas Tripathi[1]

[1] Department of Computer Science and Engineering Graphic era university, Dehradun, Uttarakhand, India,
Emails: monikapandey234@gmail.com, vishuchaudhary28@gmail.com, vikastripathi.be@gmail.com

Abstract: Automatic teller machine (ATM) is a service provided to carry out banking transactions and has become the necessity of everyday life. ATM's enables access to the bank in order to make variety of transactions such as cash withdrawal, check balance or transfer money. However, besides facilitating the banking needs, the ATM's lacks in providing the security against the several ATM frauds like money snatching and attacks on customers. In recent years there has been lot of work shown in the field of abnormal activity recognition in ATM surveillance to meet the security requirements. In this paper, we present a framework to detect the unusual activities practiced at the ATM installations. The proposed approach makes use of medial frames and combination of Zernike moments with gradient based descriptors to extract the relevant features from video. The features obtained are classified using Random forest classifier. The proposed method is successful in determining the normal and abnormal human activities with an average accuracy of 95.52%.

Keywords: Medial Frames, Abnormal Activity, ATM Frauds, Zernike Moment, Gradient Based Descriptor, Random Forest

1. INTRODUCTION

In the field of computer vision an intelligent video surveillance has become an important factor, mainly in classifying the abnormal actions in order to safeguard the events. Video surveillance in various fields has contributed to enhance security and protection [3]. Video surveillance works to detect moving object [1] [2] through an image sequence. The detection is based on motion analysis, shape analysis, object tracking, pose recognition etc [5]. ATM surveillance is the application of video surveillance. ATM is a computerized service which allows the financial transaction exempting the functions of bank clerk or teller [4]. The increase in ATM installations supports large number of transactions quickly. In the present scenario there is a tremendous increase in the fallacious activities like robbery, snatching, murder and other crimes which alarms to install an effective system to ensure the safety [7]. ATM's generally make use of CCTV cameras to monitor the activities. The CCTV cameras are not automated and thus require an authority to monitor them. Hence an automated system is required which can automatically detect the abnormal activity in the camera view [8]. Since, ATM's are installed

in public areas they also acquire physical security. The physical security in the ATM's does not lead to the ideal security system as, the intruder may harm the authority and no information of the crime is left behind. In this paper we present a method that can recognize the uncommon actions by combination of two shape descriptors. The four categories of human actions are classified under four categories (when single or multiple person in the camera view), they are: single, single abnormal, multiple and multiple abnormal, as we can see in the fig.1. Further paper is organized in the following manner. Section II, review the work done in this field. Section III describes the methodology proposed. Section IV gives the analysis and results of the proposed method and section V makes the conclusion.

2. LITERATURE REVIEW

Video based surveillance system has enhanced the security and protection in various aspects. There has been a lot of work accomplished in this field to provide secure surveillance. In this section we present the work done to improve the video surveillance. Several approaches have been presented for recognizing human actions, bobick and davis [9] in their paper make use of temporal templates using two components of templates i.e. Motion History Image (MHI) and Motion Energy Image (MEI) for human action recognition. Motion History Image represents global spatio-temporal information in the image sequences, this method is proficient for motion analysis. However, Motion History Image gives a fixed motion strength for every point in the foreground [a]. A survey by AtiqurAhad et al [10] depicts the analysis of human movements using Motion History Image and its variants. The methods based on Motion History Image gives the motion flow by using intensity of each pixel. The survey also describes various other descriptors, describing human movements by its motion, shape and other components. Descriptors like spatio-temporal interest feature points (STIP), histograms of oriented flow (HOF), histograms of oriented gradients (HOG) gives effective computation and representation of actions. In STIP features are locally analyzed, STIP detects areas having high intensity differentiation in both time and space as spatio-temporals and generally suffers from sparse spatio-temporal interest feature point detection [11]. HOG is window based descriptor that is generally computed to determine the interest point. In this window is aimed on the point of interest and splits into a grid of (n x n), further frequency histogram is generated from each cell of the grid, to show the edge orientation in the cell [12] while HOF deals in information of motion using optical flow, it directly quantizes orientation of flow vectors [13]. The use of other approaches like optical flow by Mahbub et al [14] shows motion in the form of the flow and give effective analysis. The optical flow distinguishes the displaced vector pixels from frames. In order to obtain temporal information in the form of features, Hu proposed a method called Hu Moments, which are invariant to translation, scale and rotation. Using Hu Moments several motion information were extracted which are independent of position, size and orientation [15]. The shape analysis methods like Hu Moments obtained from the motion image can easily give feature vectors for motion recognition [16]. Bobick et al [17] use hu moments for extracting features for temporal images. Hu Moments descriptor is extensively used shape descriptor as it is simple approach and less computational [18-20]. Further, descriptors like Fourier Descriptors and Zernike moments etc were also presented for describing information based on shape analysis. Fourier Descriptors [21] specify an image to fixed number of points, but this approach discriminates when image size varies due to number of points. Zernike Moments are the enhancement to Hu moments, whose magnitude is invariant to rotation [21]. Zernike moments are used in recognition of pattern i.e. as a shape descriptor in recognizing images. These moments are superior to other moment function like geometric moments, Hu moments in terms of efficiency and robustness in manifestation to noise and quantization error. The orthogonality property of Zernike moments helps to reduce the redundancy to near zero in a set of moment functions. Thus each moment corresponds to independent feature of the image.[22]. Laptive et al [23] describes histograms of optical flow and histograms of spatio-temporal gradients. They present the immovability in video for different motions. They also describes how adaptive velocity features gives motion in an unknown camera. Overall, the researchers discover that gradient and moment function based approaches gives best outputs. Also, representation based on HOG is proficient as orientation information obtained is robust

Figure 1: Different Classes of The Input Video

to changes in the camera view [24]. This approach is effectively used for local features representation and profound description of objects in the images [25, 26].

Up until various methodologies have been presented for motion recognition but, in recent papers the motion is represented by combining various motion and shape descriptors to obtain higher accuracy[5]. In our approach we have used medial frames to generate higher temporal information and further using the fusion of motion based gradients with Zernike moments. We have described that both the frameworks together can effectively determine the ATM events.

3. METHODOLOGY

The proposed methodology make use of various computer vision based techniques to detect normal and abnormal activity practiced in indoor environment like ATM room. The method consist of the camera feed in form of video, which is converted to frames. These frames are further used to extract temporal information using medial frames. The medial frame is provided to the combination of Zernike moments and gradient based descriptor to

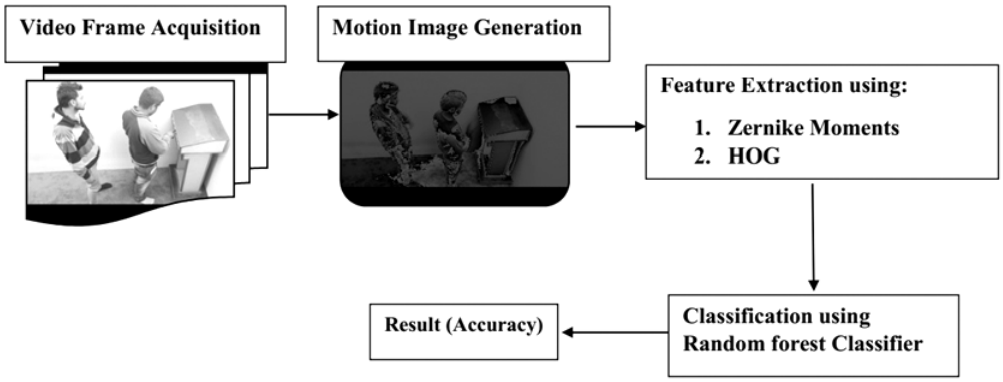

Figure 2: Proposed framework architecture

Algorithm 1
Generation of descriptor

Algorithm: Generation of descriptor

Input: Video with resolution (minimum) 320 × 240

1. Initialize x=0	1. Initialize x
2. While x < frames do	2. Frames= Number of Frames
3. Initialize y=0	3. Initialize y
4. While y < n	4. n= buffer size
5. new=new +I(x+y)	5. Addition of frames
6. y = y + 1	6. Increment y
7. Z=new/n	7. Average of added frames (new)
8. x = x+1	8. Increment x
9. Compute Zernike	9. Computing Zernike Moments
10. Compute HOG	10. Computing Histogram of Gradient

obtain motion features in the image sequence. The features thus obtained are classified using Random Forest classifier. This framework can be clearly depicted in the fig.2 showing architecture of the proposed method.

3.1. Input video

The videos provided to the algorithm presented has the minimum resolution of 320 x 240, which are recorded in the indoor environment i.e. ATM room. Videos captured are analyzed under the four categories: (i) single: when single person is in the video and performing normal activities. (ii) Single abnormal: when single person is in the video and performs abnormal activity. (iii) Multiple: when multiple person are in the camera view and act normally. (iv) Multiple abnormal: when multiple person are in the view and abnormal activities are performed.

3.2. Descriptors

Zernike moments: Zernike moments are orthogonal moments which are effective in image representation. These are rotation invariant. Zernike moments are constructed from Zernike polynomials which are orthogonally independent and hence the image representation do not suffers from overlapping or redundancy. These polynomials are defined on the unit circle, $x^2+ y^2=1$ using eq 1 and eq 2,

$$V_{nm}(x, y) = V_{nm}(\rho, \theta) = R_{nm}(\rho)e^{jm\theta} \tag{1}$$

$$R_{nm}[\rho] = \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s} \tag{2}$$

Where, n is positive integer and m is an integer such that n-|m| is even and $|m| \leq n$,

$$\rho = \sqrt{x^2 + y^2}, \text{ and}$$

$$\theta = \tan^{-1}\frac{y}{x}.$$

The Zernike moment of order n with repetition m is shown in eq 3,

$$A_{nm} = \frac{n+1}{\pi}\sum_n\sum_n f(x, y)V_{nm}(x, y),\ x^2+ y^2 \leq 1 \tag{3}$$

HOG: Hog is responsible for withdrawing shape information of object in image using intensity gradients and edge directions. Hog calculate x and y derivative of image (I) using convolution operation as shown in eq. 4 and eq. 5:

$$I_x = I * D_x \text{ Where, } D_x = [-1\ 0\ 1] \tag{4}$$

$$I_y = I * D_y \text{ Where, } D_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \tag{5}$$

Now, magnitude and gradient of I can be computed using eq. 6 and eq. 7:

$$|G| = \sqrt{I_x^2 + I_y^2} \tag{6}$$

$$\theta = arctan \frac{I_x}{I_y}. \tag{7}$$

Finally cell histograms are created and normalized using L2 normalization shown in eq. 8.

$$f = \frac{v}{\sqrt{v_2^2 +}} \tag{8}$$

Here is a small constant and represents un-normalized vector containing all histograms of the current block.

3.3. Action Classification

Random forest is a method of classification which works by creating multiple decision trees during training. Here we have trained model using random forest classifier which creates 100 trees.

The algorithm 1 represents complete working of fig.1

4. RESULTS AND ANALYSIS:

The framework has been trained and tested using python and opencv on computer having AMD A6, 2.0 Ghz processor with 4GB RAM on the videos for computing various shape descriptors. The framework has test for four classes: single, single abnormal, multiple, multiple abnormal on 49 videos (10 single, 10 single abnormal, 20 multiple and 9 multiple abnormal). We have made our own dataset of frame resolution 320 x 240 for training and testing purpose. The framework is trained using these videos for extracting features from the motion images. Testing is done on different video from the one used for training. The algorithm has tested for multiple frames provided to the descriptor.

Table 1
Accuracy of descriptor in percentage (%)

Number of frames	*Zernike*	*Fusion of Zernike with Gradient based descriptor*
1	68.7974	94.9753
5	71.0396	94.7195
10	67.4959	95.0249
15	64.9254	95.5224

Table 2
Confusion matrix of (a) Zernike and (b) Fusion of Zernike using 15 frames

	a	*b*	*c*	*d*
a = single	107	0	0	0
b = single_abnormal	0	35	0	3
c = multiple	1	0	179	10
d = multiple_abnormal	0	0	4	63

(a)

	a	*b*	*c*	*d*
a = single	107	0	0	0
b = single_abnormal	0	13	4	21
c = multiple	25	28	110	27
d = multiple_abnormal	4	20	12	31

(b)

In purposed algorithm the video is taken as input and is converted to single channeled frames. Then, firstly we extract feature of single frame at a time using shape descriptors. But, we analyzed that the single frame does not contain any temporal information which results in low accuracy. So, to extract the temporal information we create medial frame, which is the mean of multiple frames (5, 10, 15) and provide this medial frame to descriptor resulting in higher accuracy rate, which can be analyzed from table.1, shows the accuracy for input videos. We also analyzed that by combining the different shape descriptors the accuracy can be increased. Hence in our method we combine Zernike moments with histograms of oriented gradients. In table. 2 (a) the confusion matrix obtained from Zernike moment is shown and in 2 (b) the confusion matrix of combination of Zernike moments and HOG is shown. From the table it is clear that Zernikemoment along with HOG provides the better prediction as compared to individual Zernike moment descriptor.

5. CONCLUSION

In this paper we have presented an algorithm for secure surveillance at ATM, which can also be used in similar grounds. The paper presents the framework for recognizing the normal and abnormal events at ATM which is required as there is a tremendous increase in the crime rate. The algorithm accuracy differs on different number of frames used and combination of varying shape descriptors. In our method the accuracy is 94.97% for single frame and 95.56% when used with 15 frames. Lower accuracy in single frame is due to less temporal information in the frame. While on increasing the frames the accuracy rate increases as the spatial temporal information is added to the frames. In our method we present the combination Zernike moment and HOG descriptor to obtain higher accuracy rate. In future several other motion and shape descriptors can be combined to obtain better accuracy. Also, other classifiers can be used for better recognition

REFERENCES

[1] Chen, P., Chen, X., Jin, B. and Zhu, X., 2012. Online EM algorithm for background subtraction.*Procedia Engineering*, *29*, pp. 164-169.

[2] Adán, Blanco. "Carlos Roberto del and JaureguizarNuñez, Fernando and García Santos,(2011)"Bayesian Visual Surveillance, a Model for Detecting and Tracking a variable number of moving objects"." In*IEEE International Conference on Image Processing*, pp. 13-14.

[3] Sujith B. Crime Detection and Avoidance in ATM: A New Framework. International Journal of Computer Science and Information Technologies. 2014.

[4] Beyan, C. and Temizel, A., 2012. Adaptive mean-shift for automated multi object tracking.*IET computer vision*, *6*(1), pp.1-12.

[5] Pandey, M. and Tripathi, V., Recent Trends in Human Motion R ends in Human Motion R ends in Human Motion Recognition: A Survey.

[6] Tripathi, Vikas, DurgaprasadGangodkar, VivekLatta, and Ankush Mittal. "Robust abnormal event recognition via motion and shape analysis at ATM installations." *Journal of Electrical and Computer Engineering* 2015 (2015): 2.

[7] Sharma, N., 2012. Analysis of different vulnerabilities in auto teller machine transactions.*Journal of Global Research in Computer Science*, *3*(3), pp.38-40.

[8] R. S. Shirbhate, N. D. Mishra, and R. P. Pande, "Video surveillance system using motion detection: a survey," InternationalJournal Advanced Networking and Applications, vol. 3, no. 5, pp.19–22, 2012.

[9] Davis JW, Bobick AF. The representation and recognition of human movement using temporal templates. InComputer Vision and Patter Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on 1997 Jun 17 (pp. 928-934). IEEE.

[10] M. Ahad, J. Tan, H. Kim and S. Ishikawa, "Motion history image: its variants and applications", Machine Vision and Applications, vol. 23, no. 2, pp. 255-281, 2010.

[11] Chakraborty, B., Holte, M.B., Moeslund, T.B. and Gonzàlez, J., 2012. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, *116*(3), pp.396-410.

[12] Hu, R. and Collomosse, J., 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, *117*(7), pp.790-806.

[13] Wang, Heng, and CordeliaSchmid. "Action recognition with improved trajectories." In*Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551-3558. 2013.

[14] U. Mahbub, H. Imtiaz and M. A. R. Ahad, "An optical flow based approach for action recognition", In Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 646-651, 2011.

[15] M. K. Hu, "Visual pattern recognition by moment invariants", IEEE Transactions on Information Theory, vol. 8, no. 2, pp. 179-187, 1962.

[16] M. Ahad, J. Tan, H. Kim and S. Ishikawa, "Motion history image: its variants and applications", Machine Vision and Applications, vol. 23, no. 2, pp. 255-281, 2010.

[17] A. Bobick and J. Davis, "The recognition of human movement using temporal templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, 2001.

[18] A. Amato and V. D. Lecce, "Semantic classification of human behaviors in video surveillance systems",WSEAS Transactions on Computers, vol. 10, pp. 343-352, 2011.

[19] Q. Chen, R. Wu, Y. Ni, R. Huan and Z. Wang, "Research on human abnormal behavior detection and recognition in intelligent video surveillance", Journal of Computational Information Systems, vol. 9, no. 1, pp. 289-296, 2011.

[20] P. Srestasathiern and A. Yilmaz, "Planar shape representation and matching under projective transformation", Computer Vision and Image Understanding, vol. 115, no. 11, pp. 1525-1535, 2011.

[21] S. Bourennane and C. Fossati, "Comparison of shape descriptors for hand posture recognition in video", Signal, Image and Video Processing, vol. 6, no. 1, pp. 147-157, 2010.

[22] Chong, C.W., Raveendran, P. and Mukundan, R., 2003. A comparative analysis of algorithms for fast computation of Zernike moments. *Pattern Recognition*, *36*(3), pp.731-742.

[23] I. Laptev and T. Lindeberg, "Velocity adaptation of space-time interest points", In proceedings of the 17th International Conference on Pattern Recognition, Cambridge, vol. 1, pp. 52-56, 2004.

[24] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition", In International workshop on automatic face and gesture recognition, vol. 12, June 26-28, Zurich, Switzerland, pp. 296-301. 1995.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, vol. 1, pp. 886-893, 2005.

[26] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627-1645, 2010.