Efficient Motion Encoding Technique for Activity Analysis at ATM Premises

Prateek Bajaj, Monika Pandey, Vikas Tripathi and Vishal Sanserwal

Abstract Automated teller machines (ATMs) have become the predominant banking channel for the majority of customer transactions. However, despite the multitudinous advantages of ATM, it lacks in providing security measures against ATM frauds. Video surveillance is one of the prominent measures against ATM frauds. In this paper, we present an approach that can be used for activity recognition in small premises such as ATM rooms by encoding the motion in images. We have used gradient-based descriptor (HOG) to extract features from image sequences. The features obtained are classified using random forest classifier. Our employed method is successful in determining abnormal and normal human activities both in case of single and multiple personnel with an average accuracy of 97%.

1 Introduction

The goal of computer vision is to facilitate the machine to interpret the world through the process of digital signal [1]. Various technologies such as motion detection and facial recognition are based on computer vision. Automating the video surveillance with the help of computer vision to detect any suspicious activity or personnel is an effective way to the cover up some flaws in the security. Video surveillance detects moving object through a sequence of images [2, 3]. ATM surveillance is a subdomain of video surveillance. ATM crime has become one of the most prominent

P. Bajaj (✉) · M. Pandey · V. Tripathi · V. Sanserwal
Department of Computer Science and Engineering, Graphic Era University, Dehradun 248002, Uttarakhand, India
e-mail: prateekbajaj552@gmail.com

M. Pandey
e-mail: monikapandey234@gmail.com

V. Tripathi
e-mail: vikastripathi.be@gmail.com

V. Sanserwal
e-mail: vishuchaudhary28@gmail.com

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,
https://doi.org/10.1007/978-981-13-1708-8_36

issues nationwide [4] as they are at the public places and vulnerable to thefts. The usual security measure in an ATM system is CCTV which is not automated and requires authority to monitor them. The slow response time of a CCTV is a reason for its under-efficiency and adds to the vulnerability of security. Automated surveillance system detects any unusual activity in their frame view and automatically takes the desired actions [5]. In a recent survey based on video surveillance, Cucchiara [6] reported that there are various problems than hinder motion detection in non-ideal conditions. Various techniques that have been used for motion analysis using automated systems are based on the framework of temporal templates and spatiotemporal templates optical flow, background subtraction, silhouettes, histograms and several others [7–10]. In this paper, we have further extended [11] by introducing a motion encoding technique called motion identifying image (MII). In MII, we have incorporated root-mean-square of thresholded images. We have analyzed four categories of human actions which are classified from a single camera view. They are single, single abnormal, multiple and multiple abnormal. The paper is further organized in the following manner: Sect. 2 reviews the recent work; Sect. 3 describes the methodology we have used; Sect. 4 gives results and analysis; and Sect. 5 concludes the paper.

2 Literature Review

Video surveillance has contributed to the enhancement of security and protection in every possible field [12]. There are various ways to detect an activity in computer vision. In this section, we present the previous work conducted to improve video surveillance. Several approaches have been presented to recognize human actions. Davis and Bobick [13] have used temporal templates using motion history image (MHI) and motion energy image (MEI) for recognizing human activity. The temporal approaches utilize vector images where each vector points motion in the image [14]. Directional motion history image (DMHI) is an extension of MHI introduced by Ahad et al. [15, 16]. Poppe [17] has presented a detailed overview of human motion analysis using MHI and its variants. Al-Berry et al. [18], motivated by MHI, introduced a stationary wavelet-based action representation, which has been used to classify variant actions. There are various descriptors such as spatiotemporal interest feature points (STIPs), histograms of oriented flow (HOF) and histograms of oriented gradients (HOGs) which are used to compute and represent actions. Space–time interest point (STIP) detectors are extensions of 2D interest point detectors that incorporate temporal information. HOG is a window-based descriptor which is used to compute interest points. Further, the window is divided into a grid of (n * n). Frequency histogram is generated from each cell of the grid to show edge orientation in the cell [19], whereas the descriptor HOF gives information using optical flow [20]. Another descriptor named Hu moments extracts interest points based on shape, independent of position, size and orientation of the image [21], and since it is a shape descriptor, it requires comparatively less computation [22–24]. Zernike moments

descriptor is another shape descriptor which is more efficient than Hu moments [21]. Sanserwal et al. [25] in their paper have proposed algorithm in which they have used HOG descriptor, Hu moments and Zernike moments descriptor for activity detection from a single viewpoint [26] Vikas et al. proposed an approach that makes use of motion history image and Hu moments to extract features from a video. Rashwan et al. [27] proposed optical flow model with new robust data obtained from histogram of oriented gradients (HOGs) computed between two consecutive frames. But the approaches such as HOG can be highly computational [28]. Huang and Huang [29] in his paper uses look-up table along with the method of integral image to speed up HOG. Uijlings et al. [30] proposed a framework that can increase the efficiency of densely sampled HOG, HOF and MBH (motion boundary histograms) descriptors. Ryan Kennedy and Camillo J. Taylor used a method in which optical flow is calculated over triangulated images [31]. In our approach, we have used three consecutive frames to encode motion into image which is then provided to gradient-based descriptor HOG. We have described that our framework can effectively recognize ATM events.

3 Methodology

The proposed methodology makes use of computer vision-based framework to detect normal and abnormal activities in indoor premises such as ATM room. Figure 1 represents working of our framework. It shows that the method consists of the camera feed in the form of video, which is converted into threshold images. Our framework consists of two parts, conversion of an image into encoded motion using MII and conversion of encoded image into features using a descriptor. MII involves preprocessing the thresholded images using root-mean-square formula. The features thus obtained are classified using random forest classifier. The algorithmic representation of our framework is shown in Fig. 2.

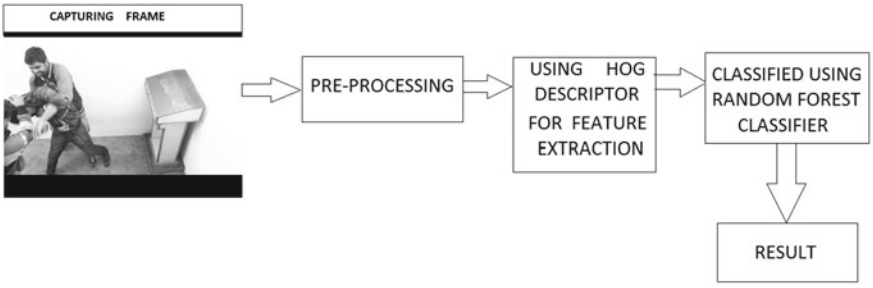

Fig. 1 Architecture for the proposed method

1. Compute Threshold images	1. Computing thereshold images
2. Initialize z=0	2. Initialize z
3. While z < frames do	3. Frames=no. Of frames
4. Initialize x=1	4. Initialize x
5. no=M(z)2+M(z+x)2	5. no=sum of square of 1st and 2nd frame
6. n=M(z)2+ M(z+x+1)2	6. n=sum of square of 1st and 3rd frame
7. $Y=\frac{\sqrt{n}+\sqrt{no}}{3}$	7.Calculating Y
8. $Y=\sqrt{Y}$	8.Calculating square root of Y
9. z = z+1	9.Increment z
10. Compute HOG	10.Computing histogram of gradient

Fig. 2 Generation of descriptor

3.1 Preprocessing

In this section, we abstract three consecutive frames and convert them into thresholded images. The method we employed for converting the frames into thresholded images is adaptive thresholding. In adaptive thresholding, we calculate different threshold values for different regions of same image. Now threshold values can be calculated using the mean of neighborhood areas or using the weighted sum of neighbor values where weights are a Gaussian window. Later, a constant is subtracted from the calculated threshold value. If the value of pixel is less than the threshold value, it is assigned to zero; otherwise, it is assigned to the desired maximum value. In our method, we have calculated threshold values for each region using mean with the block size (size of neighborhood area which is used to calculate threshold value) of eleven and the constant (which is subtracted) two. The value of constant may vary for some other set of videos. Let T(x, y) is a pixel after thresholding, t be the thresholded value, m be the maximum value that can be assigned to the pixel and I(x, y) is a pixel of a frame. The equation for adaptive thresholding is given in Eq. (1).

$$T(x, y) = \begin{cases} m & \text{if } I(x, y) \geq t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Further, we compute squares of each pixel in first, second and third frames we get after thresholding as shown in Eqs. (2), (3) and (4). Then, we calculate two values A and B, by adding the values in Eqs. (2) and (3), (2) and (4), respectively, as shown in Eq. (5) and Eq. (6). The square root of these A and B is calculated and is then

divided by the number of frames which in our case is three as depicted in Eq. (7). The operation of square root is again applied to the achieved result C, Eq. (8).

$$\mathrm{F1} = \mathrm{IMG1}^2 \tag{2}$$

$$\mathrm{F2} = \mathrm{IMG2}^2 \tag{3}$$

$$\mathrm{F3} = \mathrm{IMG3}^2 \tag{4}$$

$$\mathrm{A} = \mathrm{F1} + \mathrm{F2} \tag{5}$$

$$\mathrm{B} = \mathrm{F1} + \mathrm{F3} \tag{6}$$

$$\mathrm{C} = \frac{\sqrt{A} + \sqrt{B}}{3} \tag{7}$$

$$\mathrm{R} = \sqrt{C} \tag{8}$$

Figure 3 shows the complete diagrammatic representation of preprocessing. After preprocessing, we obtain motion identifying image which is then fed to our descriptor HOG for feature extraction.

3.2 Descriptor

We have used histogram of orientation gradient (HOG) to compute features of motion identifying image. HOG describes the appearance of a local object within an image by distribution of intensity gradient or edge directions. The image that we give as an input to the descriptor is divided into small regions, which are called cells. These cells are connected. Histogram of gradient directions is calculated for each pixel within these cells. HOG computes the derivative of image (M) with respect to x and y as shown in Eqs. 9 and 10.

$$\mathrm{Mx} = \mathrm{M} * \mathrm{DX} \text{ where } \mathrm{DX} = [-1\ 0\ -1] \tag{9}$$

$$\mathrm{My} = \mathrm{M} * \mathrm{DY} \text{ where } \mathrm{DY} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \tag{10}$$

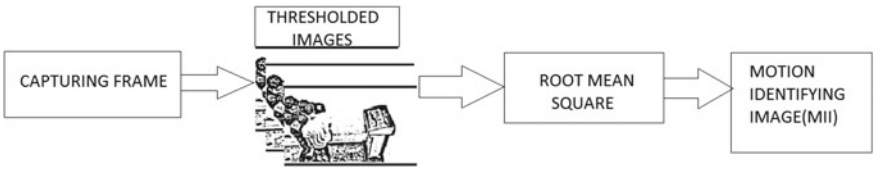

Fig. 3 Preprocessing

Further, we calculate magnitude and gradient of M in Eqs. 11 and 12.

$$|\mathrm{Gr}| = \sqrt{M_x^2 + M_y^2} \tag{11}$$

$$\emptyset = \arctan\left(\frac{M_x}{M_y}\right) \tag{12}$$

Finally, cell histograms are created and then normalized using L2 normalization as shown in Eq. 13.

$$\mathcal{F} = \frac{n}{\sqrt{n_2^2 + \partial}} \tag{13}$$

Here, n represents vector without normalization containing all histograms of current block and ∂ represent small constant.

We have used random forest classifier that works by creating multiple decision trees during training. In our case, the model had been trained using random forest classifier which creates 100 trees.

4 Results and Analysis

The proposed framework has been tested and trained, using Python 3.0 and OpenCV on computer having the specifications Intel i5 clocked at 2.4 GHz processor with the RAM of 16 GB, on videos for calculating various shape descriptors. The videos analyzed by the presented algorithms have a minimum resolution of 320 × 240. These videos are recorded in indoor premises such as ATM room. We have analyzed four categories of video captured as shown in Fig. 4: (i) single: when normal activities are being performed by a single person in a single camera view; (ii) single abnormal: when abnormal activities are being performed by a single person in a single camera view; (iii) multiple: when normal activities are being performed by a multiple person in a single camera view; (iv) multiple abnormal: when abnormal activities are being performed by a multiple person. There are a total of 49 videos in all the four classes (10 single, 10 single abnormal, 20 multiple and 9 multiple abnormal). In India, it is common for multiple personnel to enter the ATM room together. So for this sole activity we have taken a class of videos multiple. The framework is trained using these videos for extracting features from image sequences. The framework uses different videos for both testing and training purposes. The algorithm is tested for three frames, and its comparison against various other algorithms is shown in Table 1. Table 2 shows the value of W, X, Y and Z, the four classes that we have used in our dataset.

In Table 1, we have given comparative analysis with two other methods for motion encoding, which produces the best accuracy when an input of ten frames is given to the descriptor. First method uses (a) motion history image (MHI) as a descriptor;

Fig. 4 Four classes of videos

Table 1 Comparison with other descriptor (in percentage %)

Algorithm used	Result (%)
1. Combination of MHI and HU Moments	95.73
2. Combination of HOG and Zernike moments	95.02
3. Motion identifying image (MII) on thresholded images	97.24

Table 2 Confusion matrix of MII on thresholded images

	W	X	Y	Z
W = Single	571	0	0	0
X = Single Abnormal	11	179	0	2
Y = Multiple	6	0	710	26
Z = Multiple Abnormal	0	0	6	332

second method uses (b) the fusion of histogram of gradient (HOG) and Zernike moments. In general, the more frames we give to the descriptor, the more accuracy we get, as temporal information increases but even after using ten frames as an input to the descriptors used in other two algorithms, their result is comparatively less than what we acquired using MII of three frames. Hence, from the figure it is clear that our descriptor MII is better in detecting motion than MHI and fusion of HOG and Zernike.

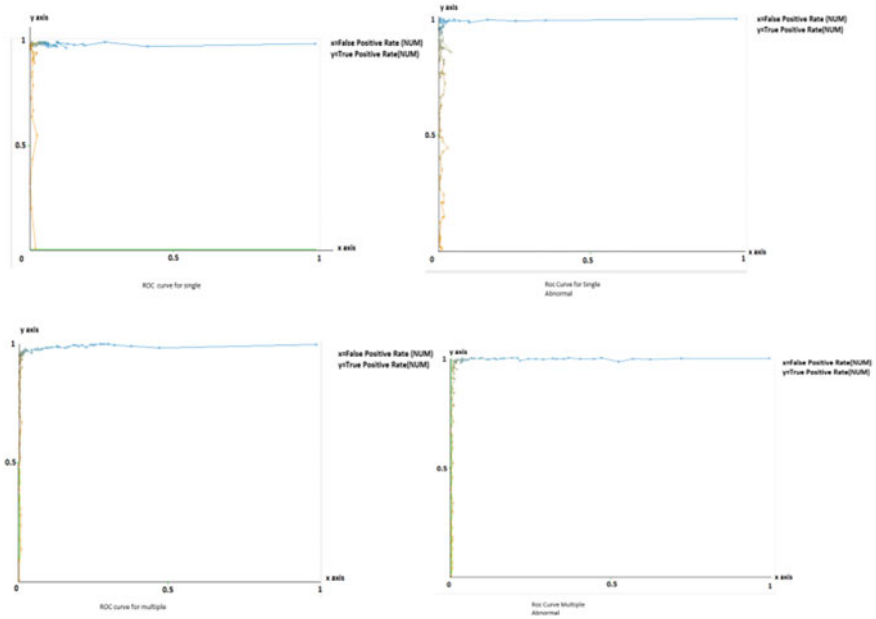

Fig. 5 ROC curve

Figure 5 shows the corresponding ROC graphs for all the four classes that is single, single abnormal, multiple and multiple abnormal for the testing dataset. In all the four graphs, the x-axis represents false-positive rate and the y-axis shows true-positive rate.

5 Conclusion

In this paper, we have proposed an algorithm that makes use of motion encoding technique called motion identifying image (MII) and a gradient-based descriptor HOG to recognize motion. It can be used in enhancing the security of ATM surveillance as well as in any other similar areas. The algorithms are tested for both normal and abnormal events with single as well as multiple personnel that can occur in ATM. It can contribute to the security of ATM as there is a tremendous increase in ATM frauds. In our method, the accuracy is about 97% when used with three frames. In the future, this motion encoding technique can be combined with any other descriptor to obtain higher accuracy. Also, an advanced and better classifier can be used for better recognition.

References

1. Wang, C., Komodakis, N., Paragios, N.: Markov random field modeling, inference & learning in computer vision & image understanding. A survey. Comput. Vis. Image Underst. **117**(11), 1610–1627 (2013)
2. Chen, P., Chen, X., Jin, B., Zhu, X.: Online EM algorithm for background subtraction. Procedia Eng. **29**, 164–169 (2012)
3. Blanco Adán, C.R., Jaureguizar, F., García, N.: Bayesian visual surveillance: a model for detecting and tracking a variable number of moving objects. In: 18th IEEE International Conference on IEEE Image Processing (ICIP), pp. 1437–1440 (2011)
4. Boateng, R.: Developing e-banking capabilities in a Ghanaian Bank. Preliminary lessons. J. Internet Bank. Commer. 213–234 (2006)
5. Kumar, P., Mittal, A., Kumar, P.: Study of robust and intelligent surveillance in visible and multi-modal framework. Informatica (Slovenia) **32**(1), 63–77 (2008)
6. Cucchiara, R.: Multimedia surveillance systems. In: Proceedings of the Third ACM International Workshop on Video Surveillance & Sensor Networks, pp. 3–10. ACM (2005)
7. Babu, R.V., Ramakrishnan, K.R.: Compressed domain human motion recognition using motion history information. In: 2003 International Conference on Image Processing, vol. 3, pp. 321–324. IEEE (2003)
8. Gupta, R., Jain, A., Rana, S.: A novel method to represent repetitive and overwriting activities in motion history images. In: 2013 International Conference on Communications and Signal Processing (ICCSP), pp. 556–560. IEEE (2013)
9. Zhou, F., De la Torre, F., Hodgins, J.K.: Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 582–596 (2013)
10. Bradski, G.R., Davis, J.: Motion segmentation and pose recognition with motion history gradients. Mach. Vis. Appl. **13**(3), 174–184 (2002)
11. Pandey, M., Sanserwal, V., Tripathi, V.: Intelligent vision based surveillance framework for ATM premises (2016)
12. Sujith, B.: Crime detection and avoidance in ATM. Int. J. Comput. Sci. Inf. Technol. 6068–6071 (2014)
13. Davis, J.W., Bobick, A.F.: The representation and recognition of human movement using temporal templates. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 928–934 (1997)
14. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marýn-Jiménez, M.J.: Automatic generation and detection of highly reliable ûducial markers under occlusion. Pattern Recognit. **47**(6), 2280–2292 (2016)
15. Ahad, M.A.R., Ogata, T., Tan, J.K., Kim, H.S., Ishikawa, S.: Directional motion history templates for low resolution motion recognition. In: 34th Annual Conference of IEEE, pp. 1875–1880 (2008)
16. Ahad, M.A.R., Ogata, T., Tan, J.K., Kim, H.S., Ishikawa, S.: Template-based human motion recognition for complex activities. IEEE International Conference, pp. 673–678 (2008)
17. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
18. Al-Berry, M.N., et al.: Action recognition using stationary wavelet-based motion images. Intelligent Systems, pp. 743–753 (2014). Springer International Publishing (2015)
19. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Comput. Vis. Image Underst. **117**(7), 790–806 (2013)
20. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
21. Hu, M.K.: Visual pattern recognition by moment invariants. IEEE Trans. Inf. Theory **8**(2), 179–187 (1962)
22. Amato, A., Lecce, V.D.: Semantic classification of human behaviors in video surveillance systems. J. WSEAS Trans. Comput. **10**, 343–352 (2011)

23. Chen, Q., Wu, R., Ni, Y., Huan, R., Wang, Z.: Research on human abnormal behavior detection and recognition in intelligent video surveillance. J. Comput. Inf. Syst. **9**(1), 289–296 (2011)
24. Srestasathiern, P., Yilmaz, A.: Planar shape representation and matching under projective transformation. Comput. Vis. Image Underst. **115**(11), 1525–1535 (2011)
25. Sanserwal, V., Pandey, M., Tripathi, V., Chan, Z.: Comparative analysis of various feature descriptors for efficient ATM surveillance framework (2017)
26. Tripathi, V., et al.: Robust abnormal event recognition via motion and shape analysis at ATM installations. J. Electr. Comput. Eng. (2015)
27. Rashwan, H.A., et al.: Illumination robust optical flow model based on histogram of oriented gradients. In: German Conference on Pattern Recognition, pp. 354–363. Springer, Berlin, Heidelberg (2013)
28. Hirabayashi, M., et al.: GPU implementations of object detection using HOG features and deformable models. In: IEEE 1st International Conference on IEEE Cyber-Physical Systems, Networks, and Applications (CPSNA), pp. 106–111 (2013)
29. Huang, C., Huang, J.: A fast HOG descriptor using lookup table and integral image (2017). arXiv:1703.06256
30. Uijlings, J., Duta, I.C., Sangineto, E., Sebe, N.: Video classification with densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off. Int. J. Multimed. Inf. Retr. **4**(1), 33–44 (2015)
31. Kennedy, R., Taylor, C.J.: Optical flow with geometric occlusion estimation and fusion of multiple frames. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 364–377. Springer International Publishing (2015)