Human Activity Recognition Using Local Motion Histogram

Awadhesh Kumar Srivastava[1,2(✉)] and K. K. Biswas[3]

[1] UTU Dehradun, Dehradun, India
srivastava_awadhesh@yahoo.co.in
[2] KIET Group of Institutions, Ghaziabad, India
[3] Bennett University, Greater Noida, India
kanad.biswas@bennett.edu.in

Abstract. Human activity recognition is an important problem in computer vision with multiple challenges. In this paper we have proposed a method for human activity recognition based on local estimation of motion in RGB videos. Background subtraction method is used on pair of consecutive frames to determine local motion, and for a small bundle of frames, the maximum magnitude of motion at a pixel is utilized to create a Projected Motion Matrix. The matrix is segmented into horizontal and vertical strips and binned histograms of each strip serve as feature descriptors. We have used these descriptors in a random forest based classification scheme and evaluated the performance on JHMDB, a publicly available human action RGB dataset.

Keywords: Human activity · Histogram · Motion projection matrix
Random forest

1 Introduction

Human activity Recognition from video has been an active area of research for more than a decade. It is a challenging problem to detect humans in video streams due to variations in pose, appearance, clothing, background clutter and illumination. Camera movement or background clutter makes it even more difficult. Potential applications include surveillance, assisted care for the elderly, monitoring of children in daycare, crowd monitoring, sports training, detection of abnormal activities and content based video retrieval. Although image based features have made considerable advances in recent years [1–4], they are not yet mature enough for many practical applications. On the other hand, most movements are characteristics of human actions, so classification accuracy can potentially be improved by paying more attention to motion information. Many researchers in this area have assumed that the camera and the background scene are essentially static. This greatly simplifies the problem because the mere presence of motion information can help us identify the action class. Towards this end, Colque et al. [13] proposed a model for capturing anomalies in human activities using orientation, velocity, and entropy. Authors calculated histogram as a feature based on optical flow. Viola *et al*. [5] point out that including motion features markedly increases the overall performance of their system.

P. Bhattacharyya et al. (Eds.): NGCT 2017, CCIS 828, pp. 908–917, 2018.
https://doi.org/10.1007/978-981-10-8660-1_69

Lot of work has been done for activity detection from RGB videos based on pose estimation and motion components. Ni et al. [6] analyzed human action through discovering the most discriminative dense trajectories group. Vrigkas et al. [7] clustered the motion trajectories and the clustered motion trajectories are used to represent human action.

Ma et al. [8] extracted video segments for partial or complete human motion. They constructed a tree based vocabulary of similar actions. Fernando et al. [9] exploited temporal ordering in videos to enumerate human actions in chronological order. They used ranking learning framework for summarization of relevant information. Zhang et al. [10] used Gaussian mixture model for modeling human action and a transfer ranking approach was used for recognizing unseen classes.

In this paper, we propose a set of features based on local estimation of *significant motion* in RGB videos. Many researchers use grid splitting of video frames for histogram calculation to generate feature descriptors [12, 14–16], To come up with features which are less sensitive to relative positioning of camera and the human in the scene, we propose to divide the motion matrix into independent horizontal and vertical strips and use the histograms of each strip as part of the feature descriptor. We show that this helps to better discern between various human actions. We use random forest classification technique as the machine learning tool, and present results on a publicly available dataset to illustrate the superiority of our method. The rest of the paper is organized as follows: Sect. 2 summarizes some related work in the area. In Sect. 3 we indicate the specific descriptors which we propose to extract from RGB video. Section 4 describes the experimental results and comparison with other state of art methods and finally Sect. 5 concludes the work.

2 Related Work

Chun and Lee [14] estimate motion flow using dense optical flow, then divide the estimations into grid cells and calculate histograms for each cell in the grid as feature descriptors. Zhang and Parker [15] proposed CoDe4D features using multi-channel orientation histogram for RGB-D data. Luo et al. [16] proposed to model the motion dynamics with robust linear dynamical systems and histograms of oriented gradients (HOG). Cheng et al. [17] proposed a framework for activity awareness using surface electromyography and accelerometer (ACC) signals. They used histogram of negative entropy to detect the starting and end point of the activity. Mukherjee et al. [18] proposed a graph theoretic technique for recognizing human actions. They used histogram of oriented optical flow and a bag-of-word approach to calculate the descriptor. Zhou and Zhang [19] proposed to encode the movements of local parts of human action. To discover elementary actions with stable states, the authors used multiple-instance formulation. Dogan et al. [20] proposed 3D volume motion templates (VMTs). To make the method view independent, the authors make a rotation with respect to a canonical orientation. Colque et al. [13] proposed Histograms of Optical Flow Orientation and Magnitude (HOFM) to detect anomalous events in videos. Tripathi et al. [12] used histogram of gradient (HOG) for detecting abnormal activity in ATM cabins.

In this paper we divide the video volume into row volumes and column volumes separately and calculate corresponding intensity histograms to reduce the sensitivity of feature vector toward the relative position of camera and object. This is described in the next section.

3 Proposed Method

RGB color frames are extracted from the action video clip, and converted to gray scale as depicted in Fig. 1a, b. The frames are bundled into small sized groups, such that each bundle contains B frames. In order to capture significant motion information, difference of consecutive frames is computed at each pixel. For each bundle, the magnitude of maximum difference at each pixel is stored in a matrix P, hereafter named as *Motion projection matrix*. Appropriate scaling of gray values at each pixel is carried out to depict the range of motion for the specific bundle. Regions having no motion will appear completely black, and areas with significant motion will appear bright gray. For a M × N frame, the difference matrix is computed using pair of consecutive frames as shown below:

$$d_i(m,n) = |f_{i+1}(m,n) - f_i(m,n)| \tag{1}$$

where m = 1, 2, …; = 1, 2, …

and i takes on values 1, 2, ….., B − 1

Next, we consider the differences at each pixel across the bundle and select the maximum d_i to create the Motion projection matrix P:

$$P(m,n) = \max(d_1(m,n),\, d_2(m,n), \ldots\ldots, d_{B-1}(m,n)) \tag{2}$$

Figure 1c depicts a typical Motion Projection Matrix.

To capture region wise movements of the video bundle, the motion projection matrix P is independently examined along both horizontal and vertical directions. Firstly P is segmented into R rows $r_1, r_2, \ldots, r_R$ where the height of each row is chosen as 5 or 10 pixels. Histogram of each horizontal strip (row) is computed and stored into 15 bins. The histogram is divided into 15 bins. Histogram of i^{th} horizontal strip is denoted by H_{r_i} which is a vector of size 15. For the R rows, we get 15 * R feature descriptors.

In a similar manner, the Motion Projection Matrix P is now segmented into C columns $c_1, c_2, \ldots, c_C$, each of width 5 or 10. Histogram of each vertical strip is computed and binned into 15 groups. H_{c_i} represents histogram of i^{th} vertical strip as depicted in Fig. 3. Each histogram H_{r_i} or H_{c_i} is in form of a vector of size 15. Correspondingly, we obtain our next set of feature descriptor with 15 * C elements.

We could have divided P into R × C grid and computed histogram of each cell. However, it turns out that cell to cell feature matching in grid splitting is highly sensitive to relative position of camera and object (i.e. if an object is performing same activity ‘near to’ or ‘away from’ camera then grid based feature matching could not perform well because the movements reside in a particular set of grids for ‘near to

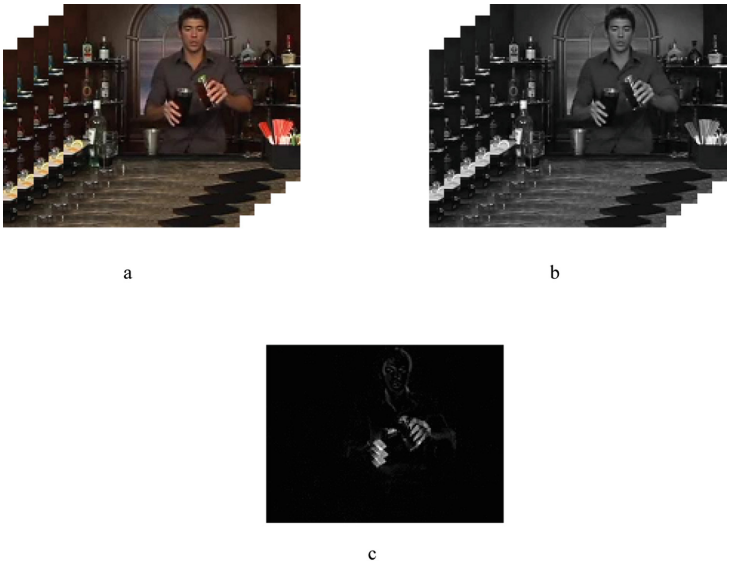

Fig. 1. (a) RGB video frames for *pour* activity, (b) gray scaled video frame with clubbing and(c) motion projection matrix for one club of frames

camera' case and other set of grids for 'away from camera' case. The same phenomenon will happen in case of left/right, up/down and various other compositions of relative positions.).

While row by row feature matching is less sensitive to the horizontal relative positions, while being more sensitive to vertical relative movements. Column by column feature matching is less sensitive to the vertical relative positions while more sensitive to horizontal relative movements of camera and object. The Proposed approach will also take care of 'near to' and 'away from' cases as shown in Fig. 2a and b.

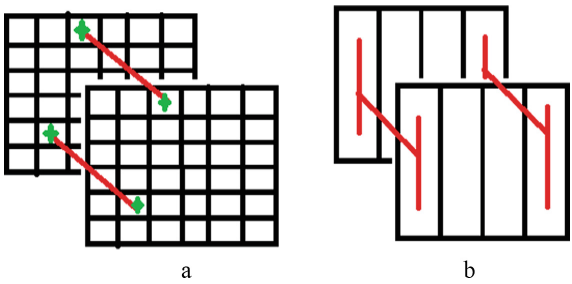

Fig. 2. Splitting of P for feature creation and matching (a) grid splitting and matching of features, (b) column splitting and matching of features

The proposed feature vector H is formed by concatenating horizontal and vertical histogram bins as shown below

$$H = [H_{r_1}, H_{r_2}, \ldots..., H_{r_R}, H_{c_1}, H_{c_2}, \ldots.., H_{c_C}] \tag{3}$$

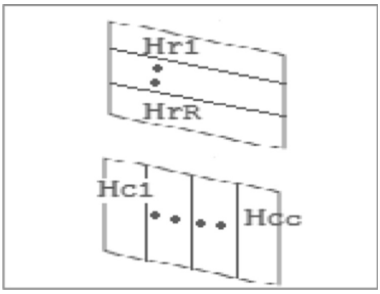

Fig. 3. Histogram calculation for horizontal and vertical regions in motion projection matrix

Algorithm 1: Feature-set Extraction

Input: video V with F frames of size NxM

1. Convert all f_i's of F from RGB color space to grayscale color space
2. Form Bundles w_1, w_2,… by clubbing B frames of F (in non overlapped fashion)
3. For every w_i
   a. Calculate motion projection matrix P_{w_i} as per equations (1) and (2)
   b. Divide P_{w_i} into R equal sized row segments
   c. Calculate histogram H_{R_i}'s as $[H_{r_1}, H_{r_2}, \ldots\ldots, H_{r_R}]$
   d. Divide P_{w_i} into C equal sized column segments
   e. Calculate histogram H_{C_i}'s as $[H_{c_1}, H_{c_2}, \ldots., H_{c_C}]$
   f. Calculate feature vector H for w_i by concatenating all histograms as equation (3)
   g. Associate activity label with H
   h. Dataset=Dataset + H :(column wise concatenation of H in to Dataset)
4. End For

Output: feature set

The size of proposed feature vector H is $15 * (C + R)$.

The output of Algorithm 1 is a set of feature vectors associated with corresponding activity labels. The outputs for various bundles are column wise concatenated to make a training dataset of features.

While the size of the various video clips in the dataset might differ for various activities, and the bundle size chosen arbitrarily, the proposed method ensures that number of feature descriptors remain fixed at 15 * (C + R) for each bundle, as it essentially extracts the histogram information.

4 Classification

Support Vector machine could be used for classification purposes. However, for large number of classes, the use of one-against-all technique creates an unbalanced dataset – usually the Positive class has very small share (5%–6%) while the negative class has the lion share (94%–95%). This may result in underperformance of the SVM algorithm because it would try to minimize the overall error. To circumvent this issue, we propose to use "random forest" for activity classification.

The method creates number of classification trees by selecting random feature vectors. The feature set is chosen randomly for training each tree in the Random Forest. Bagging is used to decrease correlation between randomly chosen trees. This makes it more immune to noise. For testing the query dataset, it is run on all the trees of the forest and the final classification is established through voting on the outcomes. We have chosen a publicly available dataset named JHMDB [11] for evaluating our proposed method.

This dataset is a joint-annotated human motion database consisting of 21 activities: (a) *brush hair,* (b) *catch*, (c) *clap*, (d) *climb stairs*, (e) *golf,* (f) *jump*, (g) *kick ball*, (h) *pick*, (i) *pour*, (j) *pull-up,* (k) *push*, (l) *run,* (m) *shoot ball*, (n) *shoot bow*, (o) *shoot gun*, (p) *sit*, (q) *stand*, (r) *swing baseball*, (s) *throw*, (t) *walk*, and (u) *wave*. The dataset consists of 36–55 clips per action class with each clip containing 15–40 frames. There are 31,838 annotated frames in total. Figure 4 illustrate some of the activities of JHMDB dataset, columns of the figure representing *catch, jump, wave* and *push* activity respectively. While first row of the figure shows the 10^{th} frame, the second row shows the 20^{th} frame of the corresponding activities.

Fig. 4. Frames of *catch*, *jump*, *wave* and *push* (column wise) activity of JHMDB dataset. First row is 10^{th} frames and second row is 20^{th} frames of the corresponding activities.

We have performed experiments on the JHMDB dataset for different values of B (number of frames in a bundle), R (number of horizontal strips in matrix P), C (number of vertical strips in matrix P) and number of trees in random forest. Typically R = C

has been chosen in our experiments. Overall Classification accuracy of various experiments are shown in Table 1.

Table 1. Classification accuracy on various parameters of the experiments for JHMDB dataset

Experiment number	Number of frames in a bundle (B)	Number of trees in Random forest	Number of rows or column segment in a motion projection matrix (R or C)	Accuracy
1	5	100	10	51.75
2	7	100	10	51.43
3	5	50	10	51.1
4	3	50	10	50.16
5	3	100	5	49.52
6	3	50	5	49.52
7	7	100	5	48.89

In the rest of the paper our discussion is based on the values of the parameters chosen as in experiment 1 of Table 1.

JHMDB dataset is a challenging dataset, since scenes are taken from movies, youtube channel etc. without imposing any constrains on light illumination, camera movements, object orientation and relative position of object with camera.

The proposed method performed well for many activities while some activities are classified with low accuracies. *Pull-up* and *shoot bow* activities are classified with accuracy of 89% and 83%. There are four more activities which have been classified with accuracy more than 70%. Activities of *sit, run, walk* and *kick ball* are classified with lower accuracies. The reason for low accuracies can be attributed to high similarity amongst some of the activities resulting in greater number of misclassifications. The proposed method's overall classification accuracy comes out to 51.75%.

Activity wise classifier accuracy is shown in Fig. 5 and the confusion matrix for activity recognition on JHMDB dataset is shown in Fig. 6.

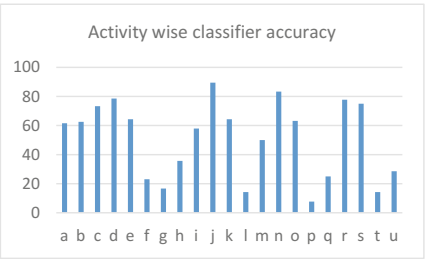

Fig. 5. Activity wise classifier accuracy for JHMBD dataset

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	0	p	q	r	s	t	u
a	0.62			0.08					0.08	0.08					0.08			0.08			
b		0.63			0.06	0.13								0.06			0.06		0.06		
c		0.07	0.73	0.07					0.07					0.07							
d			0.07	0.79		0.07					0.07										
e				0.07	0.64					0.07					0.14				0.07		
f	0.15	0.15			0.08	0.23	0.08	0.08			0.08			0.08				0.08			
g	0.08	0.08			0.08	0.17	0.17		0.08	0.08					0.08			0.08	0.08		
h						0.07		0.36		0.07	0.14	0.07		0.14		0.07			0.07		
i			0.11		0.05				0.58				0.05	0.05	0.05			0.05			0.05
j									0.11	0.89											
k			0.07	0.07						0.07	0.64	0.07		0.07							
l				0.14		0.14		0.07			0.07	0.14		0.07	0.07		0.14	0.07		0.07	
m	0.14	0.14		0.07				0.07			0.07		0.50								
n	0.06										0.06			0.83	0.06						
o					0.05					0.05	0.16			0.11	0.63						
p			0.23								0.15			0.08	0.15	0.08	0.08	0.15		0.08	
q	0.08		0.08	0.08							0.00	0.08			0.08	0.17	0.25	0.08	0.08		
r			0.06						0.06	0.06	0.06							0.78			
s				0.06											0.06		0.06	0.06	0.75		
t				0.14	0.14						0.29			0.07	0.14			0.07		0.14	
u			0.07						0.07		0.21	0.07		0.07	0.07			0.07		0.07	0.29

Fig. 6. Confusion matrix for JHMDB dataset

Results of proposed method are compared with other state of art techniques in Table 2. It can be seen that our approach is performing better than all the histogram and trajectory based approaches of Jhuang et al. [11].

Table 2. Comparison with other methods.

Sr No	Author	Approach	Feature			
			Traj.	HOG	HOF	L. M. Histogram
1	Jhuang et al. [11]	(1) baseline	40.0	32.9	40.1	
		(2) of pmask	38.5	31.9	46.0	
		(3) pf pmask	36.4	32.8	48.0	
		(4) pf Dmask	38.0	32.2	46.4	
		(5) pf pmask of outside pmask	43.0	36.1	44.1	
		(6) (4) + (5)	46.2	35.2	51.7	
		(7) bbox F w. [21]	37.7	33.9	39.0	
		(8) bbox F	38.5	34.9	42.2	
		(9) bboxIm	42.7	46.9	44.5	
		(10) DmaskIm	41.4	47.0	45.6	
2	Proposed	Row and column wise splitting				**51.75**

5 Conclusion

In this paper we have proposed a method for human action recognition based on local estimation of motion in RGB videos. Background subtraction method is used on pair of consecutive frames to determine local motion, and for a small bundle of frames, the maximum magnitude of motion at a pixel is saved to create a Projected Motion Matrix. The matrix is segmented into horizontal and vertical strips and binned histograms of each strip serve as feature descriptors. We have used these descriptors in a random forest based classification scheme and evaluated the performance on a publicly available human action RGB dataset.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: a review. ACM Comput. Surv. **43**, 16 (2011)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, San Diego, California, USA, pp. 886–893 (2005)
3. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, San Diego, California, USA, pp. 876–885, June 2005
4. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 69–82. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24670-1_6
5. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: Proceedings of the 9th International Conference on Computer Vision, Nice, France, vol. 1, pp. 734–741 (2003)
6. Ni, B., Moulin, P., Yang, X., Yan, S.: Motion part regularization: improving action recognition via trajectory group selection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3698–3706 (2015)
7. Vrigkas, M., Karavasilis, V., Nikou, C., Kakadiaris, I.A.: Matching mixtures of curves for human action recognition. Comput. Vis. Image Understand. **119**, 27–40 (2014). https://doi.org/10.1016/j.cviu.2013.11.007
8. Ma, S., Sigal, L., Sclaroff, S.: Space-time tree ensemble for action recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5024–5032 (2015)
9. Fernando, B., Gavves, E., Oramas, J.M., Ghodrati, A., Tuytelaars, T.: Modeling video evolution for action recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp 5378–5387 (2015)
10. Zhang, Z., Wang, C., Xiao, B., Zhou, W., Liu, S.: Robust relative attributes for human action recognition. Pattern Anal. Appl. **18**, 157–171 (2015). https://doi.org/10.1007/s10044-013-0349-3
11. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.: Towards understanding action recognition. In: ICCV (2013)
12. Tripathi, V., Mittal, A., Gangodkar, D., Kanth, V.: Real time security framework for detecting abnormal events at ATM installations. J. Real Time Image Process. (JRTIP), 1–11 (2016). https://doi.org/10.1007/s11554-016-0573-3

13. Colque, R.V.H.M., Caetano, C., de Andrade, M.T.L., Schwartz, W.R.: Histograms of optical flow orientation and magnitude to detect anomalous events in videos. IEEE Trans. Circ. Syst. Video Technol. **27**(3), 673–682 (2017). https://doi.org/10.1109/TCSVT.2016.2637778
14. Chun, S., Lee, C.-S.: Human action recognition using histogram of motion intensity and direction from multiple views. IET Comput. Vis. **10**, 250–257 (2016)
15. Zhang, H., Parker, L.E.: CoDe4D: color-depth local spatio-temporal features for human activity recognition from RGB-D videos. IEEE Trans. Circ. Syst. Video Technol. **26**(3), 541–555 (2016)
16. Luo, G., Yang, S., Tian, G., Yuan, C., Hu, W., Maybank, S.J.: Learning human actions by combining global dynamics and local appearance. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **36**(12), 2466–2482 (2014)
17. Cheng, J., Chen, X., Shen, M.: A framework for daily activity monitoring and fall detection based on surface electromyography and accelerometer signals. IEEE J. Biomed. Health Inf. **17**(1), 38–45 (2013)
18. Mukherjee, S., Biswas, S.K., Mukherjee, D.P.: Recognizing human action at a distance in video by key poses. IEEE Trans. Circ. Syst. Video Technol. **21**(9), 1228–1241 (2011)
19. Zhou, W., Zhang, Z.: Human action recognition with multiple-instance Markov model. IEEE Trans. Inf. Forensics and Secur. **9**(10), 1581–1591 (2014)
20. Dogan, E., Eren, G., Wolf, C., Baskurt, A.: Activity recognition with volume motion templates and histograms of 3d gradients. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 4421–4425 (2015)
21. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 168–181. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15567-3_13