Human Action Recognition Based on Spatio-temporal Features

Nikhil Sawant and K.K. Biswas

Dept. of CSE, IIT Delhi-110016, India
{mcs072899|kkb}@cse.iitd.ac.in
http://cse.iitd.ac.in/~mcs072899

Abstract. This paper studies the technique of human action recognition using spatio-temporal features. We concentrate on the motion and the shape patterns produced by different actions for action recognition. The motion patterns generated by the actions are captured by the optical flows. The Shape information is obtained by Viola-Jones features. Spatial features comprises of motion and shape information from a single frame. Spatio-temporal descriptor patterns are formed to improve the accuracy over spatial features. Adaboost learns and classifies the descriptor patterns. We report the accuracy of our system on a standard Weizmann dataset.

1 Introduction

Human action recognition is becoming increasingly important for automation of video analysis. With the growing need of surveillance related applications the research in the field of action recognition has been fueled in past few years. A detailed survey of action recognition techniques has been presented by Gavrila [1]. The researchers have used space-time features to identify the specific action [2,3]. Computer vision scientists have tried motion based techniques [4,5,6] as any action is associated with some motion. Niu et. al. [4] have used both motion and eigen shape features to carry out view invariant activity recognition. Bag of words [5] is recently has been used for the task of action recognition.

We make use of both shape and motion patterns as well as space-time pattern features. Adaboost learns and detects the patterns of different actions. Patterns are spatio-temporal features made up of motion and shape information. The paper is organized as follows: Section 2 explains about target localization. In sections 3 and 4, we discuss the motion and shape descriptors respectively followed by discussion on spatio-temporal features in section 5. The learning process is explained in section 6 and we present our results and conclusions in section 7.

2 Target Localization

Target localization helps reducing the search space. Background subtraction is used to generate a silhouette of the target as shown in Figure 1(b). We assume

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 357–362, 2009.

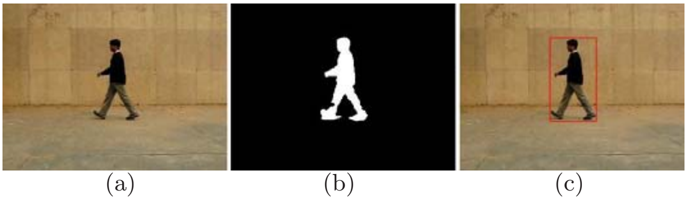


Fig. 1. Target localization. (a) shows a video frame, (b) is the extracted silhouette where the background is stable, (c) is the original frame (a) with *ROI* marked with the help of silhouette in (b).

that the action is being performed in front of a stable background. With the help of silhouette information the region of Interest (*ROI*) can be determined as shown in Figure 1(c). Once the *ROI* is marked we can concentrate only on the area inside *ROI*.

3 Motion Descriptor

It has been shown that different activities produce different motion patterns. We make use of Lucas - Kanade method [7] to generate the optical flows in the *ROI* to capture the motion patterns. The advantage of this method is that it comparatively yields robust and dense optical flow fields.

Organizing optical flows using averaging. After computation of optical flow our job is to arrange it in some fashion so that a descriptor can be formed.

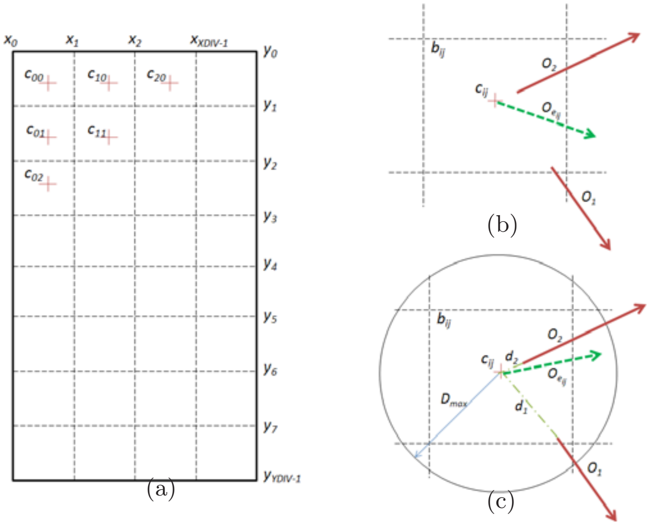


Fig. 2. (a) shows a grid overlaid over the *ROI*, (b) Simple averaging: $O_{e_{ij}}$ formed by averaging of all the vectors within the box b_{ij} i.e O_1 and O_2, (c) Weighted average: O_2 has more influence on $O_{e_{ij}}$ compare to O_1

A fixed size grid is overlaid over the ROI as shown in Figure 2(a). This grid divides the silhouette into boxes $\{b_{ij}, ...\}$ with centres at $\{c_{ij}, ...\}$ respectively. An intuitive way is to average out the optical flow from each box to form the effective optical flow $O_{e_{ij}}$ which is formulated in Eq 1.

$$O_{e_{ij}} = \frac{\sum_{k=1}^{m'} O_k}{\sum_{k=1}^{m'} 1} \quad \text{where } (x_i \le x_{O_k} < x_{i+1}) \;\&\; (y_j \le y_{O_k} < y_{j+1}) \quad \forall k \tag{1}$$

Here m' is the set of optical flows within the box b_{ij}. $O_{e_{ij}}$ is the effective optical flow for box b_{ij}. x_{O_k} and y_{O_k} are the x and y co-ordinates of the k^{th} optical flow O_k respectively. In Figure 2(b), $O_{e_{ij}}$ has equal contribution from all the optical flows present within the box.

A possible drawback in the simple averaging method is that all the optical flows have same weight irrespective of their distance from the centre of the box. Thus the net optical flow may be swayed by an optical flow in different direction sitting at the boundary of the box. Thus we present a new weighted method to compute the net optical flow for each box.

Organizing optical flows using weighted average. As shown in the Figure 2(c) we sum up the contribution of various optical flow vectors at c_{ij} the centre of each grid cell. We assume that only the flow vectors lying within distance D_{max} from the grid center would be allowed to influence the computation of the effective optical flow. The contribution of each flow is weighted inversely by the distance from the center. The net flow is computed by the following Eq 2.

$$\begin{gathered} O_{e_{ij}} = \frac{\sum_{k=1}^{m''} ((D_{max} - d_k) O_k)}{\sum_{k=1}^{m''} (D_{max} - d_k)} \quad \text{where } D_{max} \ge d_k \quad \forall k \\ \text{and } d_k = \sqrt{(x_{O_k} - x_{c_{ij}})^2 + (y_{O_k} - y_{c_{ij}})^2} \end{gathered} \tag{2}$$

Here m'' is the set of optical flows within range of D_{max} from the centre c_{ij}. $x_{c_{ij}}$ and $y_{c_{ij}}$ is nothing but the x and y co-ordinates of the centre c_{ij} of the box b_{ij} respectively.

Figure 3(c) shows optical flow after applying weighted average on each box b_{ij} of the overlaid grid. In order to form the descriptor the effective optical flow $O_{e_{ij}}$ is split into $O_{ex_{ij}}$ and $O_{ey_{ij}}$ in the respective direction. For a grid size of $(M$x$N)$, we have $2MN$ values representing the motion descriptor.

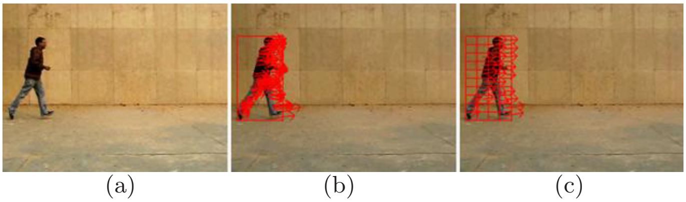


Fig. 3. Organized optical flows. (a) original frame, (b) unorganized optical flows, (c) organized optical flows.

4 Shape Descriptor

The shape of a silhouette is also a characteristic of the action being performed. Shape information is helpful when motion information is noisy or not sufficient. Niu et. al. [4] used shape information by adding the value of each and every pixel to the descriptor, but this requires resizing the silhouette.

Differential shape information. We propose use of rectangular features introduced by Viola-Jones [8] for face detection. These features are simple and have already proven their usefulness. Figure 4(b)(c) show $two - rectangle$ and $four - rectangle$ features used by us. Rectangular features are applied at box level and not at pixel as done by Viola-Jones. We overlaid the grid over the ROI as shown in Figure 4(a). Each box is assigned the percentage of foreground pixels within.

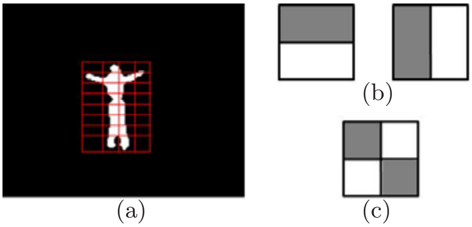


Fig. 4. Shape Information. (a) silhouette with the grid overlaied within the ROI, (b) $two - rectangle$ features, (c) $four - rectangle$ features.

5 Spatio-temporal Descriptors

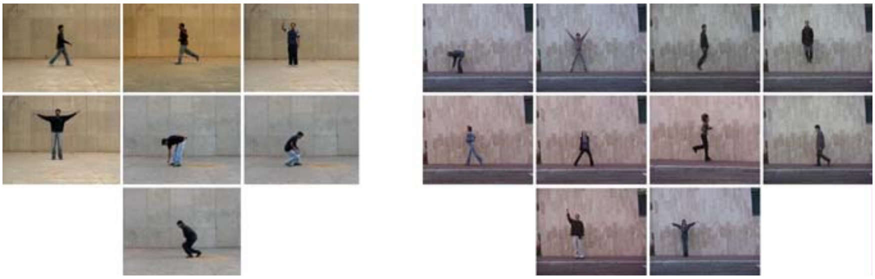
The motion and the shape descriptor described in previous two sections can be obtained from each individual frame of the video sequence. In order to improve the accuracy of action recognition we make use of spatio-temporal features. Spatio-temporal features not only carry information about current frame but also about neighboring frames. We fix the number of frames for all the videos under consideration, and call it $TLEN$. To reduce the computational overhead in considering all the frames, we select the frames with a fixed offset called $TSPAN$, for example if we choose $TSPAN$ as 3, we pick up every third frame from the video clip and stack all these together to extract the spatio-temporal features. *TSPAN* helps in reducing the descriptor length without much change in the accuracy.

6 Learning with Adaboost

We use standard Adaboost learning algorithm [9]. Adaboost is the state of art boosting algorithm. Linear decision stumps are used as the weak hypothesis. We train our system for patterns of all the chosen actions. Once trained the System can recognize the patterns produced by different actions. Our training and testing data is mutually exclusive. Also we have used different subjects for training and testing sequences.

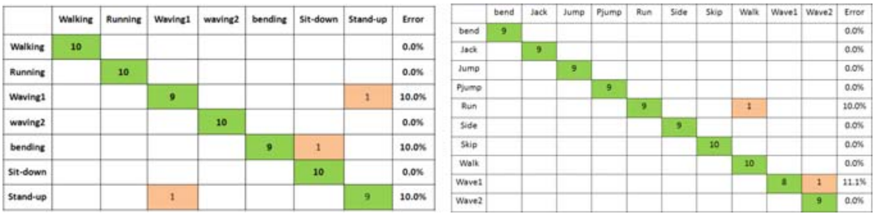
7 Results and Conclusion

We conducted our initial experiments on a small dataset built by us, infront of stable background. The dataset has 7 actions performed by 5 to 8 actors. Our dataset contains around 10 videos of each action. Figure 5(a) shows the snapshot of our dataset, various actions covered are walking, running, waving1, waving2, bending, sit-down, stand-up (left to right, top to bottom). We also tested our method for standard Weizmann dataset [3] . Snapshot of the Weizmann dataset is shown in Figure 5(b), various actions covered are bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2 (left to right, top to bottom), each action has 9 videos performed by 9 separate actors. The fixed grid of 4 x 8 is overlaid on the *ROI*. For spatio temporal features, we experimented with *TLEN* as 5 and *TSPAN* as 5. Figure 6(a) shows the confusion matrix for our dataset. As we see there is 0% error rate for walking, running, waving2, sit-down. Overall error rate of recognition is 4:28%. For standard Weizmann dataset, results are shown in Figure 6(b). There is slight error in run and wave1 actions, rest all the actions are performed with 0% error. Our error is rate 2:17% which is better than 16:3% reported recently by T. Goodhart [5].



(a) (b)

Fig. 5. Dataset used. (a) Our own dataset consisting of 7 actions, (b) standard Weizmann Dataset containing 10 actions.



	Walking	Running	Waving1	waving2	bending	Sit-down	Stand-up	Error
Walking	10							0.0%
Running		10						0.0%
Waving1			9				1	10.0%
waving2				10				0.0%
bending					9	1		10.0%
Sit-down						10		0.0%
Stand-up			1				9	10.0%

(a)

	bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2	Error
bend	9										0.0%
Jack		9									0.0%
Jump			9								0.0%
Pjump				9							0.0%
Run					9			1			10.0%
Side						9					0.0%
Skip							10				0.0%
Walk								10			0.0%
Wave1									8	1	11.1%
Wave2										9	0.0%

(b)

Fig. 6. Confusion matrix. (a) Results on our dataset, (b) Results on standard Weizmann Dataset.

Conclusion. We propose a method for action recognition technique which uses motion and shape features. Spatio-temporal patterns generated by different actions clearly highlight the differences between them. Results of our technique are better than some of the recently reported results [5]. We have successfully shown that spatio-temporal features consisting of motion and shape patterns can be used for action recognition with stable background.

References

1. Gavrila, D.M.: The visual analysis of human movement: a survey. Comput. Vis. Image Underst. 73(1), 82–98 (1999)
2. Sullivan, J., Carlsson, S.: Recognizing and tracking human action. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 629–644. Springer, Heidelberg (2002)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV 2005: Proceedings of the Tenth IEEE International Conference on Computer Vision, Washington, DC, USA, pp. 1395–1402. IEEE Computer Society, Los Alamitos (2005)
4. Niu, F., Abdel-Mottaleb, M.: View-invariant human activity recognition based on shape and motion features, pp. 546–556 (December 2004)
5. Goodhart, T., Yan, P., Shah, M.: Action recognition using spatio-temporal regularity based features, 745–748 (31 2008 - April 4 2008)
6. Danafar, S., Gheissari, N.: Action recognition for surveillance applications using optic flow and SVM. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 457–466. Springer, Heidelberg (2007)
7. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision (1981)
8. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2), 137–154 (2004)
9. Freund, Y., Schapire, R.: A short introduction to boosting. J. Japan. Soc. for Artif. Intel. 14(5), 771–780 (1999)