Chapter 10
Action Recognition Based on Hierarchical Model

Yang-yang Wang, Yang Liu and Jin Xu

Abstract The feature representation of human actions is one of the important factors which influence the recognition accuracy of actions. Usually the recognition accuracy is higher, when the feature simultaneously includes both appearance and motion information. However the dimensions of the feature space is high, and this leads to high computational cost. To overcome this problem, we propose a hierarchical model for action recognition. In the first hierarchy, we adopt box features to divide the actions into two classes, according to whether or not legs are all almost stayed in a static place. In the second hierarchy, we construct different structure of motion feature descriptors to represent different kinds of actions, and use nearest neighbor classifier to obtain the final classification results. Experiments on the Weizmann dataset demonstrate the effectiveness of the proposed method.

Keywords Action recognition · Box feature · Hierarchical model · Motion feature

10.1 Introduction

Human action recognition is widely applied in the fields of surveillance, image and video retrieval and so on. Recently many researches have been done for the recognition, however, it still is a challenging problem due to diverse variations of human body motions, for example, occlusion, scale, individual appearance and individual motion fashion. To achieve good recognition performance, a good representation with rich appearance and motion information is of vital. Usually if a feature includes not only still appearance but also dynamic motion information, the

Y. Wang (✉) · Y. Liu · J. Xu
College of Automation, Shenyang Aerospace University, Shenyang,
People's Republic of China
e-mail: wyy2004101@yahoo.com.cn

E. Qi et al. (eds.), *The 19th International Conference on Industrial Engineering and Engineering Management*, DOI: 10.1007/978-3-642-38391-5_10,
© Springer-Verlag Berlin Heidelberg 2013

feature dimension is high, and it is more discriminative to classify the actions. But this leads to high computational costs. To overcome this problem, we propose a hierarchical model for action recognition using multi-features, and show its performance compared to other exist algorithms.

The rest of this paper is organized as follows. Section 10.2 describes related work. Section 10.3 elaborates the details of our model. Section 10.4 reports the experiment results. Conclusion is given in Sect. 10.5.

10.2 Related Work

Many works have been proposed for action recognition, and some reviews (Moeslund et al. 2006; Poppe 2010; Ji and Liu 2010) have provided a detailed description of the action recognition framework. Here, we only focus on the action recognition that has related to our work.

Features based on shape or appearance is traditional representation in videos analysis. This kind of representation is usually related to global properties of human. First human is separated from the background, which is called box. And then all kinds of descriptors based on silhouette or edges are described. Deng et al. (2010) compute the silhouettes of the human body, and extracts points with equal interval along the silhouette. Through constructing the 3D DAISY descriptor of each point, the space–time shape of an action is obtained. Bobick and Davis (2001) also make use of the silhouette, but they employ motion energy images (MEI) and motion history image (MHI) to describe the actions. Ji and Liu (2009) present contour shaper feature to describe multi-view silhouette images. In Nater et al. (2010) a signed distance transform is applied to the box with fixed size. In Weinland and Boyer (2008) silhouette templates are used to match. But the recognition rate may be reflected by the accurate degree of the localization of the box, and the effect of background subtraction.

Another popular feature representation is about motion information. Messing et al. (2009) use the velocity histories of tracked key points to recognize the actions. Laptev et al. (2008) first use HOF descriptor to represent local motion information. Recently the combination of shape and motion features has received more attention. Lin et al. (2009) construct a shape-motion descriptor, and model a shape-motion prototype tree using hierarchical k-means clustering. Klaser et al. (2008) make use of HOG3D descriptor to combine motion and shape information simultaneously. Although these descriptors include more rich information, and the recognition accuracy is improved, the dimensions of the descriptors are bigger than those of the descriptors with single shape or motion information. A trade-off between computation complexity and accuracy should be considered. Therefore, we propose a hierarchical model, in the first stage, a coarse classification is made according to box features, and then different descriptors are designed to different class, this can reduce the dim of the feature vectors.

10.3 Proposed Method

There are five main steps to our approach: first, preprocessing video to achieve the human box; second, extracting global box features; third, preliminary dividing the actions into two different classes based on box features; forth, computing respective motion features for the actions which belong to different classes; and fifth, recognizing actions using nearest neighbor classifier and voting algorithm. In the following, we describe each step in turn.

10.3.1 Preprocess

We start from a video of k frames which are described in a RGB space. For feature extraction and recognition, we use background subtraction and filtering to segment the foreground object, which is the box of a person in each frame of an action sequence. In our paper, the motion direction of the same action is all aligned to the same direction, e.g. for the action ‘running’, the direction is appointed from left to right, Fig. 10.1b, c. Furthermore, the resulting silhouettes are converted to a binary representation. Examples of the box are illustrated in Fig. 10.1, and the representation of the boxes of a video S is

$$S = \{B_1, B_2, \ldots\ldots, B_k\}. \tag{10.1}$$

10.3.2 Coarse Classification Based on Box Feature

For each box B_i ($i = 1,\ldots, k$), a three-dimensional box feature vector is computed, i.e. $F_{bi} = [cx_i, cy_i, r_i]$, here cx_i, cy_i, r_i denote the coordinates of the center, aspect ratio of the box. According to the variance of coordinates and aspect ratio in an image sequences, we can easily divided the simple actions into two classes. Class 1 includes the actions which the motions mainly focus on the leg movements, such as jogging, running, walking, skipping and so on. And the actions which legs are all almost stayed in a static place are all belong to the Class 2, such as bending, waving.

10.3.3 Local Weighted-Motion Feature

Having classified the actions into two classes, we can construct different weighted-motion feature according to their particular characteristic.

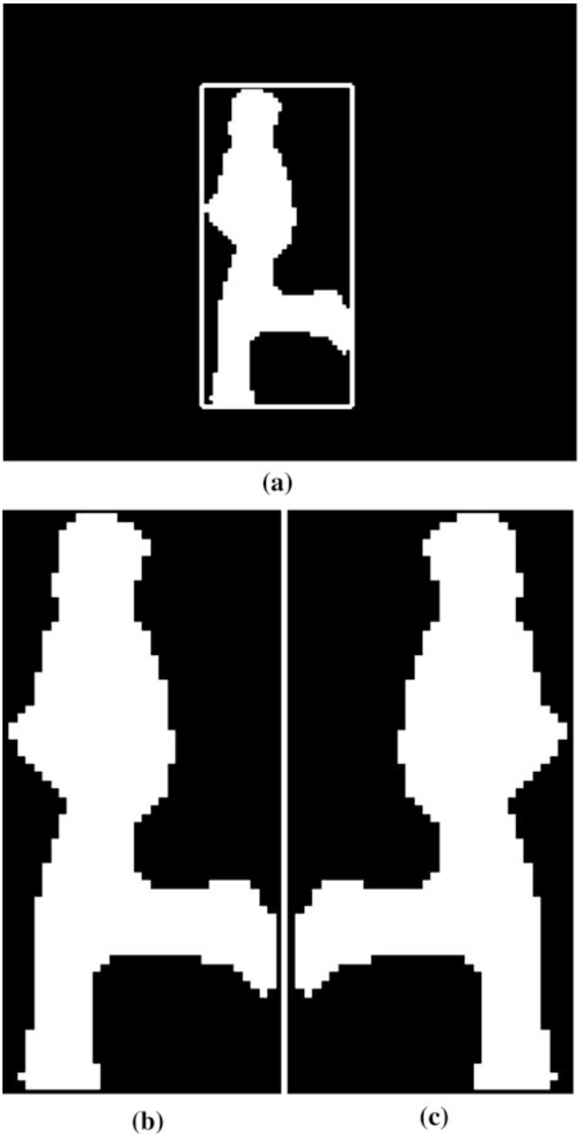


Fig. 10.1 Example of the box: **a** a frame after background subtraction, **b** extracting the box, **c** aligned the motion direction of the box

(1) *Box normalization*. The size and the position of the box in each frame is different, after getting the box feature F_{bi}, we normalized the boxes to equal size (80 × 80 in our case) while maintaining aspect ratio.

(2) *The division of subregions*. When a person is in a state of motion in a video sequence, his body is continually varying in time series. For the actions which belong to Class 1, the movements of its upper limb and lower limb are more important. Whereas for the actions in Class 2, e.g. running, the variances in subregions of legs are more important. In order to exactly describe the

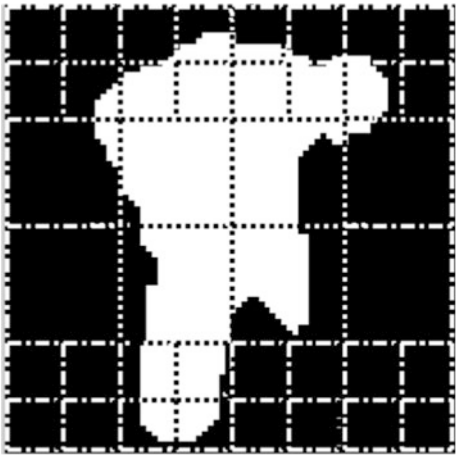


Fig. 10.2 The division of subregions for Class 1

variation, the different division of subregions is adopted to different action classes. First, for all kinds of actions, each normalized box is divided into *n* square subregions. And then for Class 1, on the basis of the division mentioned above, for the first and last row of the box, we use a further division, as shown in Fig. 10.2. And for Class 2, for the second and third row of the box, each subregion is divided into 4 square grids, as shown in Fig. 10.3.

(3) *Weighted-motion feature representation*. Optical flow can be made to describe the human motion. We calculate optical flow using Horn and Schunck algorithm (Horn and Schunck 1981). To each pixel in each subregion, the velocity of optical flow $V\ (x, y, V_x, V_y)$ is computed, where (x, y) represents the image location of V, and (V_x, V_y) are the x and y components of the velocity V. Besides, for reducing the influences of noise dynamic backgrounds, we use a Gauss filter to smooth the subregions before calculating optical flow. According to the (V_x, V_y), the magnitude and orientation of the V is computed

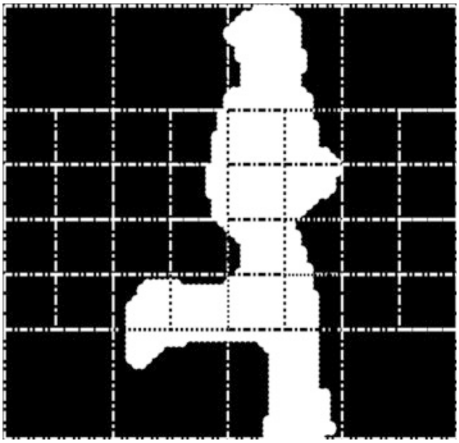


Fig. 10.3 The division of subregions for Class 2

$$M\left(V_x, V_y\right) = \left|\left(V_x, V_y\right)\right|, \tag{10.2}$$

$$O\left(V_x, V_y\right) = \arctan\left(V_y / V_x\right). \tag{10.3}$$

$O(V_x, V_y)$ is divided into eight equal sectors in polar coordinates, and the $M(V_x, V_y)$ can be regarded as the weights. The weighted-motion histogram's x-axis reflects the eight orientations. The histogram's y-axis shows the contribution within each orientation. And for each subregion we achieve a histogram.

Further, the subregions with grids are more important than others, in order to precisely represent actions, for each grid in a subregion a single motion histogram is computed. Therefore, the number of histogram is $8 + 8 \times 4 = 40$.

10.3.4 Final Classification Based on Weighted-Motion Feature

Having got the features of each frame of an image sequence, we respectively classify two different classes of actions. Although the features of respective class are different, the classifier is the same.

NNC (Nearest Neighbor Classifier) is used to recognize the action categories. For the fth frame of an unknown video Su, we get a distance vector $D_f = [d_1, d_2, \ldots, d_q]$, here d_j ($j = 1, \ldots, q$) denotes the distance between the features of the fth frame in Su and the features of each frame in training set, q denotes the number of all of frames in the training set. The minimum distance in D_f is picked up, and the label of the corresponding training frame is assigned to the fth frame of Su. And finally each frame of Su gets a label; the final action label of Su is determined by voting algorithm.

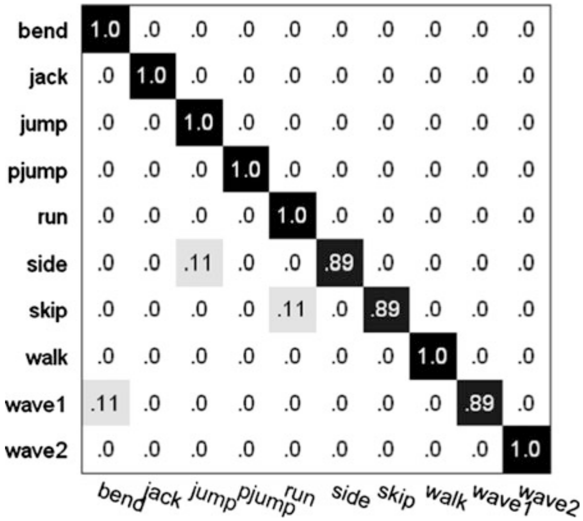
10.4 Experiments

To evaluate the performance of our algorithm, the popular benchmark dataset—Weizmann dataset is used in action recognition. The Weizmann dataset (Blank et al. 2005) consists of 90 videos of nine actors performing ten different actions. These actions are walking, running, jumping, siding, and bending, one-hand waving, two-hands waving, jacking, jumping in place and skipping. Evaluations were done with a leave-one-out cross-validation.

Our recognition result is shown in Table 10.1 and compared with Lin' method (Lin et al. 2009) and Scovanner's method (Scovanner et al. 2007). In addition, the confusion matrix is shown in Fig. 10.4. The data on the diagonal line present the numbers of the correct identification, and other data in the graph are the numbers of the misclassification. The average recognition rate of our approach is 96.7 %.

Table 10.1 Results using different features on Weizmann dataset

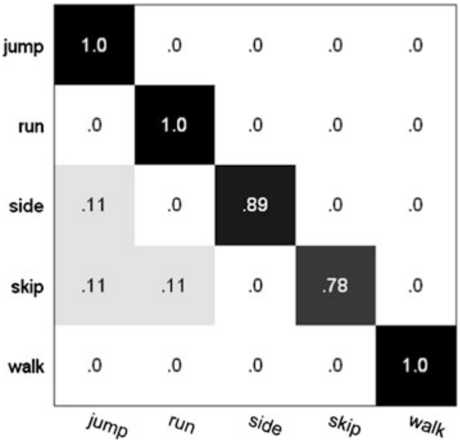
Descriptor	Average precision (%)
Shape-motion tree (Lin et al. 2009)	94.44
3D SIFT (Scovanner et al. 2007)	82.6
Ours	96.7



	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
jack	.0	1.0	.0	.0	.0	.0	.0	.0	.0	.0
jump	.0	.0	1.0	.0	.0	.0	.0	.0	.0	.0
pjump	.0	.0	.0	1.0	.0	.0	.0	.0	.0	.0
run	.0	.0	.0	.0	1.0	.0	.0	.0	.0	.0
side	.0	.0	.11	.0	.0	.89	.0	.0	.0	.0
skip	.0	.0	.0	.0	.11	.0	.89	.0	.0	.0
walk	.0	.0	.0	.0	.0	.0	.0	1.0	.0	.0
wave1	.11	.0	.0	.0	.0	.0	.0	.0	.89	.0
wave2	.0	.0	.0	.0	.0	.0	.0	.0	.0	1.0

Fig. 10.4 Confusion matrix for Weizmann dataset

Besides we exchange the motion feature structure of the two classes. When the actions in Class 2 use the feature structure of Class 1, as shown in Fig. 10.5, the average accuracy is 93.4 %. When the actions in Class 1 use the feature structure of Class 2, the influence of recognition rate is very small. The results prove that our method is effective.



	jump	run	side	skip	walk
jump	1.0	.0	.0	.0	.0
run	.0	1.0	.0	.0	.0
side	.11	.0	.89	.0	.0
skip	.11	.11	.0	.78	.0
walk	.0	.0	.0	.0	1.0

Fig. 10.5 Confusion matrix for Class 2. The form of the feature representation is used Class 1's

10.5 Conclusion

We have presented a novel hierarchical feature representation. First, the actions are divided into two classes based on the box feature. And then for the different classes of actions, the different motion feature structure is designed. This method is clearly useful for action recognition. However background segmentation and the alignment of boxes are required in the method. One future research direction would be to improve the robustness of feature descriptors in scenes where boxes can not be obtained reliably.

References

Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) In: Proceedings of 10th IEEE conference on computer vision, ICCV'05, Beijing, China, pp 1395–1402

Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267

Deng C, Cao X, Liu H, Chen J (2010) A global spatio-temporal representation for action recognition. In: Proceedings of 20th conference on pattern recognition, ICPR'10, Istanbul, Turkey, pp 1816–1819

Horn BKP, Schunck BG (1981) Determining optical flow. Artif Intell 17(1–3):185–203

Ji X, Liu H (2009) View-invariant human action recognition using exemplar-based hidden markov models. In: Proceedings of 2nd conference on intelligent robotics and applications, ICIRA'09, Singapore, pp 78–89

Ji X, Liu H (2010) Advances in view-invariant human motion analysis: a review. IEEE Trans Syst Man Cybern Part C Appl Rev 40(1):13–24

Klaser A, Marszalek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of 19th British machine vision conference, BMVC'08, Leeds, UK, pp 995–1004

Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proceedings of 21st IEEE conference on computer vision and pattern recognition, CVPR'08, Anchorage, AK, pp 1–8

Lin Z, Jiang Z, Davis LS (2009) Recognizing actions by shape-motion prototype trees. In: Proceedings of 12th IEEE conference on computer vision, ICCV'09, Kyoto, Japan, pp 444–451

Messing R, Pal C, Kautz H (2009) Activity recognition using the velocity histories of tracked keypoints. In: Proceedings of 12th IEEE conference on computer vision, ICCV'09, Kyoto, Japan, pp 104–111

Moeslund TB, Hilton A, Kruer V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Underst 104(2–3):90–126

Nater F, Grabner H, Van Gool L (2010) Exploiting simple hierarchies for unsupervised human behavior analysis. In: Proceedings of 23rd IEEE conference on computer vision and pattern recognition, CVPR'10, San Francisco, CA, pp 2014–2021

Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990

Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of 15th conference on multimedia, MM'07, Augsburg, Germany, pp 357–360

Weinland D, Boyer E (2008) Action recognition using exemplar-based embedding. In: Proceedings of 21st IEEE conference on computer vision and pattern recognition, CVPR'08, Anchorage, AK, pp 1–7