Human Action Recognition Using Distribution of Oriented Rectangular Patches

Nazlı İkizler and Pınar Duygulu

Dept. of Computer Engineering, Bilkent University, Ankara, Turkey
{inazli,duygulu}@cs.bilkent.edu.tr

Abstract. We describe a "bag-of-rectangles" method for representing and recognizing human actions in videos. In this method, each human pose in an action sequence is represented by oriented rectangular patches extracted over the whole body. Then, spatial oriented histograms are formed to represent the distribution of these rectangular patches. In order to carry the information from the spatial domain described by the bag-of-rectangles descriptor to temporal domain for recognition of the actions, four different methods are proposed. These are namely, (i) frame by frame voting, which recognizes the actions by matching the descriptors of each frame, (ii) global histogramming, which extends the idea of Motion Energy Image proposed by Bobick and Davis by rectangular patches, (iii) a classifier based approach using SVMs, and (iv) adaptation of Dynamic Time Warping on the temporal representation of the descriptor. The detailed experiments are carried out on the action dataset of Blank et. al. High success rates (100%) prove that with a very simple and compact representation, we can achieve robust recognition of human actions, compared to complex representations.

1 Introduction

Understanding human motion is one of the appealing, yet challenging problems of computer vision. Reliable and effective solutions to this problem can serve many areas, ranging from human-computer interaction to security surveillance. However, although tracking is now a usable technology, understanding what people are doing is still at its infancy.

Human action recognition has been a widely studied topic (for extensive reviews see [12,8]). Yet, the solutions to the problem are very premature and very specific to dataset at hand.

For human motion understanding in videos, there are three major approaches: First, one can use temporal logics to represent crucial order relations between states that constrain activities. Examples to such approaches include Pinhanez and Bobick [19,20], describing a method based on interval algebra, and Siskind [24] describing methods to infer activities related to objects using a form of logical inference.

Second, one can use spatio-temporal templates to identify instances of activities. Spatio-temporal patterns date at least to Polana and Nelson [21]. Thinking

A. Elgammal et al. (Eds.): Human Motion 2007, LNCS 4814, pp. 271–284, 2007.

actions as such spatio-temporal templates is made famous by Bobick and Davis [2]. They introduce Motion-Energy-Image and Motion-History-Image templates for recognizing different motions. Efros *et al.* [5] use a motion descriptor based on optical flow of a spatio-temporal volume. Blank *et al.* [1] also define actions as space-time shapes, making use of Poisson distributions to define the details of such shapes.

Third general approach to recognize human motion is to use models of dynamics, such as hidden markov models ([3], [27], [18]), conditional random fields [26], finite state models [11,10]. These (mostly generative) models rely on modeling the details of the action dynamics and need lots of training data to build effective models. Ikizler and Forsyth [13] show how to make use of motion capture data in such a case.

We argue that, human pose encapsulates many useful clues for recognizing the ongoing activity. Actions can mostly be represented by the configuration of the body parts, before building complex models for understanding the dynamics. Using this idea, we focus on building a pose descriptor which can be used to discriminate actions. Unlike most of the methods that use complex modeling of the body configurations, we follow the analogy of Forsyth *et al.* [7], which represents the body as a set of rectangles, and explore the layout of these rectangles.

Our pose descriptor is based on a basic intuition: human body can be represented by a collection of oriented rectangles in the spatial domain and the orientations of these rectangles form a signature for each action. Rather than detecting and learning the exact configuration of body parts, we are only interested in the distribution of the rectangular regions which may be the candidates for the body parts.

This idea follows from the bag-of-words approach, where the images are represented by collection of regions, ignoring their spatial relationships. Bag-of-words approach – which is adapted from text retrieval literature – has shown to be successful for object and scene recognition [6,25] and for annotation and retrieval of large image and video collections [16,28]. In such approaches, the images are represented by the distribution of words from a fixed visual vocabulary (i.e. image patches) which is usually obtained by vector quantization of visual features.

Histogramming is an old trick that has been frequently used in computer vision research. For action recognition, Freeman and Roth [9] used orientation histograms for hand gesture recognition. Recently, Dalal and Triggs used histograms of oriented gradients (HOGs) for human detection in images [4], which is shown to be quite successful.

Our main contribution is to adapt the bag-of-words approach for action recognition, by considering the distribution of higher level rectangular patches which are candidates for body parts.

In the following, we first describe our "bag-of-rectangles" pose descriptor, which represents the human figures as a distribution of oriented rectangular patches. Then, we utilize four different methods to recognize the actions. These are namely; frame by frame voting, global histogramming, SVM classification

and Dynamic Time Warping. The detailed experiments are carried out on the data set of Blank *et al.* [1].

2 Bag-of-Rectangles Method

Following the body plan analogy of Forsyth *et al.* [7], we represent human body as a collection of rectangular patches and we base our motion understanding approach on the fact that orientations and positions of these rectangles change over time w.r.t. actions carried out. With this intuition, our algorithm first extracts rectangular patches over the human figure available in each frame, and then forms a spatial histogram of these rectangles by grouping over orientations. We then evaluate the changes of these histograms over time.

More specifically, given the video, first, the tracker identifies the location of the subject. Then, the bounding box around its silhouette is extracted. This bounding box is then divided into a $N \times N$ equal-sized spatial bins. While forming these spatial bins, the ratio between the body parts, i.e. head torso and legs, is taken into account. At each time t, a pose is represented with a histogram H_t formed based on the orientations of the rectangles in each spatial bin. This process is depicted in Fig. 1.

Having formed the spatio-temporal rectangle histograms for each video, we match any newly seen sequence to the examples at hand and label the videos accordingly. We now describe the steps of our method in greater detail.

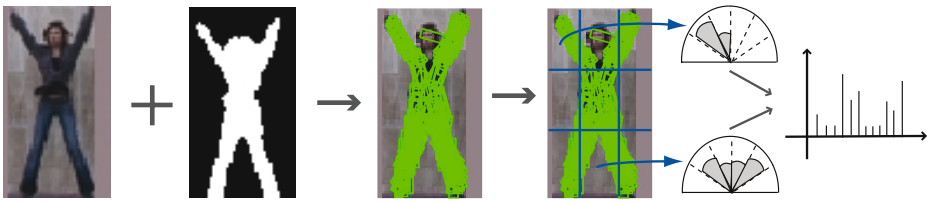


Fig. 1. Here, feature extraction stage of our approach is shown (this figure is best viewed in color). First, the human figure in each frame is extracted using background subtraction or an appropriate tracker. Using these silhouettes, we search for the rectangular patches that can be candidates of limbs. We do not discriminate between legs and arms here. Then, we divide the bounding box around the silhouette into an equal-sized grid and compute the histograms of the oriented rectangles inside each region. We form our feature vector by combining the histograms coming from each subregion.

2.1 Rectangle Extraction

For describing the human pose, we make use of rectangular patches. These patches are extracted in the following way:

1) The tracker fires a response for the human figure. This is usually done using a foreground-background discrimination method. The simplest approach is to apply background subtraction, after forming a dependable model of the

background. The reader is referred to [8] for a detailed overview of the subject. In our experiments, we use the results of a background subtraction scheme (the extracted masks by Blank *et al.*) to localize the subject in motion. Note that any other method that extracts the silhouette of the subject will work just fine.

2) We then search for rectangular regions over the human silhouette using convolution of a rectangular filter on different orientations and scales. We make use of undirected rectangular filters, following Ramanan *et al.* [22]. The search is performed using 12 tilting angles, which are 15° apart, covering a search space of 180°. Note that since we don't have the directional information of these rectangle patches, orientations do not cover 360° but its half. To tolerate the difference in the limb sizes and varying camera distances to the subject, we perform the rectangle convolution over multiple scales.

More formally, we form a zero-padded rectangular Gaussian filter G_{rect} and produce the rectangular regions $R(x, y)$ by means of the convolution of the binary silhouette image $I(x, y)$ with this rectangle filter G_{rect}.

$$R(x, y) = G_{rect}(x, y) \circ I(x, y) \tag{1}$$

where G_{rect} is zero-padded rectangular patch of a 2-D Gaussian $G(x, y)$

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{2}$$

Higher response areas to this filter are more likely to include patches of particular kind. The filters used are shown in Fig. 2.

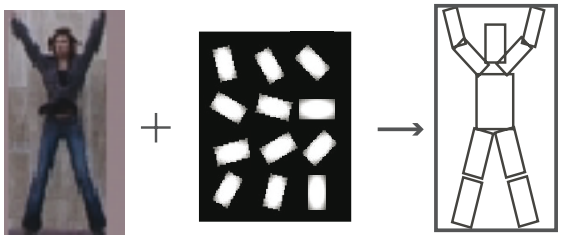


Fig. 2. The rectangular filtering process is shown. We use zero-padded Gaussian filters with 15° tilted orientations over the human silhouette. We search over various scales, without discriminating between different body parts. The perfect rectangular search for the given human subject would result in the tree structure to the right.

To tolerate noise and imperfect silhouette extraction, this rectangle search allows a portion of the candidate regions to remain non-responsive to the filters. Regions that have low overall responses are eliminated this way. Then, we select k of the remaining candidate regions of each scale by random sampling (we used $k = 300$).

One can also perform a special search for the torso rectangle, which is considerably larger than limb rectangles and omit this torso region while searching

for the remaining body parts and then form rectangular histograms. Example images of this rectangular search is given in Fig. 3. In (a) torso is excluded, in (b) the rectangles are searched over the whole silhouette. We evaluate the effects of such a torso exclusion step in the experiments section.

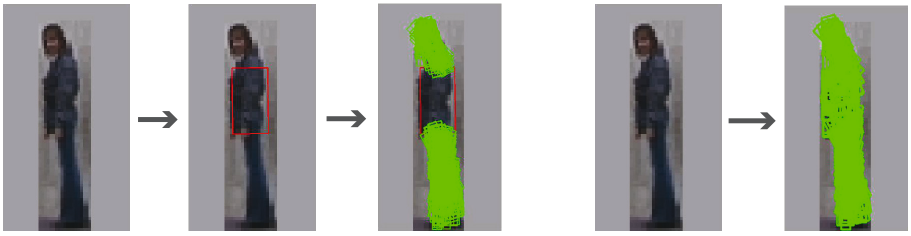


Fig. 3. Rectangle detection with and without torso detection (best viewed in color). In (a) first, the torso region is detected. This is done by applying a larger rectangular filter and taking the mean of the responses. After finding the torso, the remaining silhouette is examined for remaining limb rectangles. In (b) the whole silhouette is used in the rectangle extraction phase.

Rectangle extraction phase results in $\sim$ 1000 rectangles per frame. While forming the histogram, one can use all of these rectangles or select representative rectangles for each limb. The rectangles which cover the silhouette as much as possible and have high responses to rectangular filters are considered as the representative ones. To achieve these constraints, the higher response candidates that are more than a specified distance apart from each other are selected. By this way, rectangle count is reduced to $\sim$ 10 rectangles per frame. Figure 4 shows this process. Although reducing rectangles gives a more compact representation, it supresses valuable information about the distribution density of the rectangles, making the approach more prone to noise. The experiments (Sect. 4) show the outcomes of such an elimination.

2.2 Pose Descriptor - Histograms of Oriented Rectangles

After finding the rectangular regions of the human body, in order to define the pose, we propose a simple pose descriptor, which is the Histogram of Oriented Rectangles (HOR). We calculate the histogram of extracted rectangular patches based on their orientations. The rectangles are histogrammed over 15° orientations resulting in 12 circular bins. In order to incorporate spatial information of the human body, we evaluate these circular histograms within a $N \times N$ grid placed over the whole body. Our experiments show that $N = 3$ gives the best results. We form this grid by splitting the silhouette over the y-dimension based on the length of the legs. The area covering the silhouette is divided into equal-sized bins from bottom to up and left to right (see Fig. 5 for details). Note that,

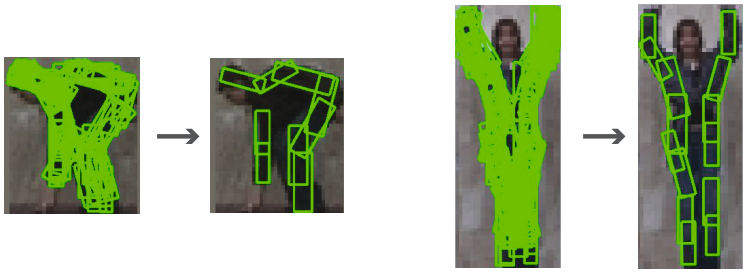


Fig. 4. Rectangle elimination process is shown. Left is a frame from the `bend` action and right is a frame from the `two-hands-wave` action. Rectangles maximizing the coverage area of the silhouette and the rectangular filter response are selected as representatives, eliminating the rest. By this way, rectangle count can be reduced to acquire a more compact representation of the body.

by this way, we let some space to the top part of the head, to allow action space for the arms (for actions like reaching, waving, etc.).

We have also evaluated effect of using 30° orientation bins and 2×2 grid, which have more concise feature representations, but coarser detail of the human pose. We show the corresponding results in Sect. 4.

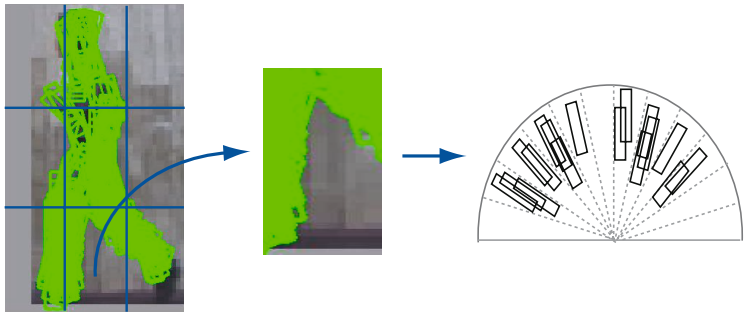


Fig. 5. Details of histogram of oriented rectangles (HORs). The bounding box around the human figure is divided into an $N \times N$ grid (in this case, 3×3) and the HOR from a single spatial bin is magnified. The resulting feature vector is a concatenation of the HORs from each spatial bin.

3 Recognizing Actions with Bag-of-Rectangles

After calculating the pose descriptors for each frame, we perform action classification in a supervised manner. There are four methods we tried in order to evaluate the performance of our pose descriptor.

3.1 Frame by Frame Voting

The simplest scheme we utilize is to perform matching based on single frames, ignoring dynamics of the sequence. That is, for each test instance frame, we

find the closest frame in the training set and employ a voting throughout the sequence. The distance between frames is calculated using Chi-square distance between the histograms (as in [14]). Each frame with the histogram H_i is labelled with the class of the frame having histogram H_j that has the smallest distance χ^2 such that

$$\chi^2(H_i, H_j) = \frac{1}{2} \sum_n \frac{(H_i(n) - H_j(n))^2}{H_i(n) + H_j(n)} \tag{3}$$

We should note that both χ^2 and L_2 distance functions are very prone to noise, because a slight shift of the human center may cause in the different binning of the rectangles, and therefore, large fluctuations in distance. One can utilize Earth Mover's Distance [23] or Diffusion Distance [15] which are shown to be more efficient for histogram comparison in the presence of such shifts by taking the distances between bins into account.

3.2 Global Histogramming

Global histogramming is similar to the Motion Energy Image (MEI) proposed by Bobick and Davis [2]. In this method, we sum up all spatial histograms of oriented rectangles through the whole sequence and form a single compact representation for the entire video. This is simply done by collapsing all time information into single dimension by summing the histograms and forming a global histogram H_{global} such that

$$H_{global}(d) = \sum_t H(d, t) \tag{4}$$

for each dimension d of the histogram. Each test instance's H_{global} is compared to that of the training instances using χ^2 distance and the closest match's label is reported. The corresponding global images are shown in Fig. 6.

3.3 SVM Classification

We also evaluate the performance of SVM-based classification with our pose-descriptor. We trained separate SVM classifiers for each action. These SVM classifiers are formed using *RBF* kernels over snippets of frames using a windowing approach. A grid search over the parameter space of the SVM classifiers is done and the best classifiers are selected using 10-fold cross validation. In our windowing approach, the sequence is segmented into k-length chunks with some overlapping ratio o, then these chunks are classified separately (we achieved the best results with $k = 15$, and $o = 3$). The whole sequence is then labelled with the most frequent action class among its chunks.

3.4 Dynamic Time Warping

Since the periods of the actions are not uniform, comparing sequences is not straightforward. In the case of human actions, the same action can be performed

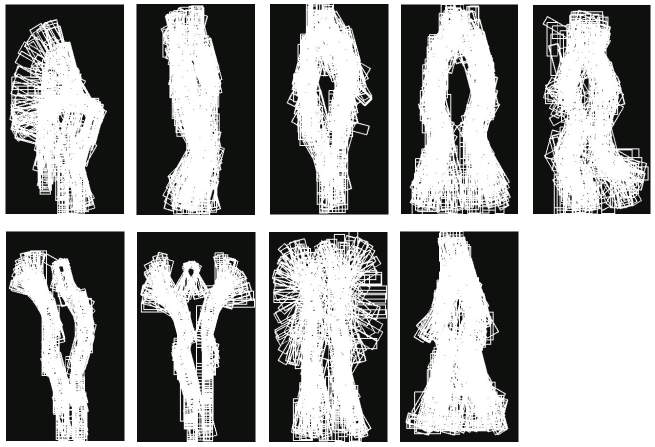


Fig. 6. Global histograms are generated by summing up all the sequence and forming the spatial histograms of oriented rectangles from these global images. In this figure, global images after the extraction of the rectangular patches are shown for 9 separate action classes.**Top row:** `bend`, `jump`, `jump in place`, `gallop sideways` and `run` actions. **Bottom row:** `one-hand wave`, `two-hands wave`, `jumpjack` and `walk` actions. These images resemble to Motion Energy Images introduced by [2], however we do not use these shapes. Instead, we form the global spatial histogram of the oriented rectangles as our feature vector.

in different speeds, resulting the sequence to be expanded or shrinked in time. In order to eliminate such effects of different speeds and to perform robust comparison, the sequences need to be aligned.

Dynamic time warping (DTW) is a method to compare two time series which may be different in length. DTW operates by trying to find the optimal alignment between two time series by means of dynamic programming. Figure 7 shows an example of time warping process. The time axes are warped in such a way that samples of the corresponding points are aligned.

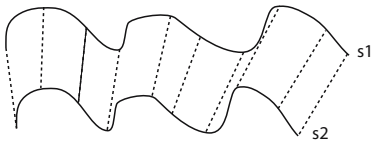


Fig. 7. Dynamic Time Warping (DTW) process in 1-d time series: The distance between corresponding sample points of two sequences $s1$ and $s2$ are calculated using dynamic programming and the time axes are warped in such a way that the corresponding sample points are aligned

More specifically, given two time series $x_1 \dots x_n$ and $y_1 \dots y_m$, the distance $D(i, j)$ is calculated with

$$D(i,j) = \left\{ \begin{array}{l} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{array} \right\} + d(x_i, y_j) \tag{5}$$

where $d(.,.)$ is the local distance function specific to application. In our implementation, we have chosen $d(.,.)$ as the χ^2 distance function as in Equation 3.

We use dynamic time warping along each dimension of the histograms separately. As shown in Fig. 8, we take each 1-d series of the histogram bins of the test video X and compute the DTW distance $D(X(d), Y(d))$ to the corresponding 1-d series of the training instance Y. We then sum up the distances of all dimensions to compute the global DTW distance (D_{global}) between the videos. We label the test video with the label of the training instance that has the smallest D_{global} such that,

$$D_{global}(X,Y) = \sum_{d=1}^{M} D(X(d), Y(d)) \tag{6}$$

where M is the total number of bins in the histograms. While doing this, we exclude the top k of the distances to reduce the effect of noise introduced by shifted bins and inaccurate rectangle regions. We choose k based on the size of the feature vector such that $k = \lfloor \#num\_bins/2 \rfloor$ where $\#num\_bins$ is the total number of bins of the spatial grid.

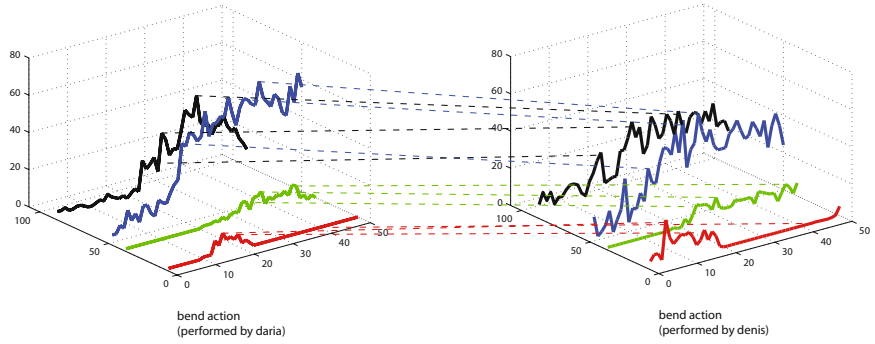


Fig. 8. Dynamic Time Warping (DTW) over 2D histograms: We calculate DTW distances between the histograms by evaluating DTW cost over single dimensions separately and summing up all costs to get a global distance between sequences. Here, histograms of two bend actions performed by different actors are shown. We try to align these sequences along each histogram dimension by DTW and report sum of the smallest distances. Note that, separate alignment of each histogram bin also allows us to handle the fluctuations in distinct body part speeds.

4 Experimental Results

4.1 Dataset

For experimental evaluation we use dataset that Blank *et al.* introduced in [1]. We used the same set of actions as [1], which is a set of 9 actions: walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place and jumping jack. We used extracted masks provided to localize the human figures in each image. These masks have been obtained using background subtraction. We test the effectiveness of our method using leave-one-out cross validation.

4.2 Rectangle Elimination

In Table 1, the effect of of rectangle elimination is shown. We observe that using the extracted rectangles as is – without selecting representative ones – give more accurate action recognition. Although we have a more compact representation with the representative rectangles, the probability distribution of the limb locations is more accurately estimated when we use all samples of rectangles in the histogram calculation.

Table 1. The accuracies of the matching methods with respect to rectangle elimination, with 15° angular bins over 3×3 grid. Although with eliminated rectangles we have a more sparse representation of the body, rectangles wrongly extracted become more significant in that case. Using all extracted rectangles gives a more robust estimation about where the actual body parts are, since body part regions are likely to produce more rectangles, resulting in denser rectangular regions.

Matching Method	Eliminated	All
FrameVoting	0.6173	0.9630
GlobalHist	0.9506	0.9630
SVM	0.9383	0.9506
DTW	0.9877	**1.0000**

4.3 Torso Detection

We can make a separate search for the torso, omit this region and form our pose descriptors based only on the candidate limb locations. In Table 2, we show the effect of torso detection on the overall accuracies. We observe that with frame voting and global histogramming methods, torso detection and exclusion helps, however, SVM and DTW classifiers suffer from slight performance degradation.

4.4 Granularity of Angular Bins

We also evaluated the choice of orientation angles when forming the histogram. Table 3 shows the results using 15° angular bins versus 30° bins. Our results indicate that there is a slight loss of information when we go from fine level orientations (i.e. 15° bins) to a coarser level (30°).

Table 2. The accuracies of the matching methods with respect to torso detection. The results presented here are over the eliminated rectangles with 15° angular bins and 3×3 grid. Note that torso detection can be useful in the case of FrameVoting and GlobalHist methods whereas SVM and DTW methods suffer from a slight performance loss.

Matching Method	No Torso	With Torso
FrameVoting	0.6173	0.6790
GlobalHist	0.9506	0.9630
SVM	0.9383	0.9259
DTW	0.9877	0.9506

Table 3. The accuracies of the matching methods with respect to angular bins. The original rectangle search is done with 15° tilted rectangular filters. To form 30° histograms, we group rectangles that fall into the same angular bins. These results demonstrate that as we move from fine to coarser scale of angles, there is a slight loss of information and thus 30° HORs become less discriminative than 15° HORs.

Matching Method	15°	30°
FrameVoting	0.9630	0.9506
GlobalHist	0.9630	0.9383
SVM	0.9506	0.9383
DTW	**1.0000**	0.9506

4.5 Grid Size

When forming the histograms of oriented rectangles, we place an $N \times N$ grid over the silhouette of the subject and form spatial histograms for each grid region. The choice of N effects the size of the feature vector (thus execution time of the matching), and the level of detail of the descriptor. Table 4 compares using 2×2 grid versus 3×3 grid. One can try further levels of partitioning, even form pyramids of these partitions. However, too dense partitioning will not make sense, since the subregions should be large enough to contain rectangle patches. Our results over this dataset indicate that, 3×3 gives better performance compared to 2×2. However, if execution time is crucial, choice of $N = 2$ will still work to a certain degree of performance.

4.6 Overall Evaluation and Comparison

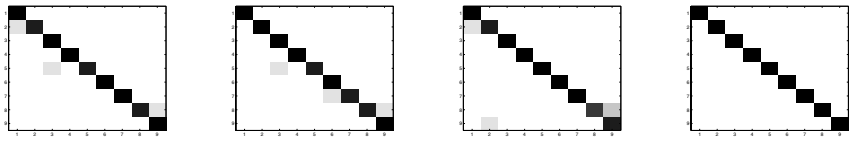
Overall, we achieve the best results with DTW matching. This is not surprising, because the subjects do not perform actions with uniform speeds and lengths. Thus, the sequences need aligning. DTW matching accomplishes this alignment over the bins of the histogram separately, making alignment of limb movements also possible. Action speed differences between body parts are handled this way.

We reach a perfect accuracy (100%) over Blank action dataset, using all extracted rectangles and the torso region with 15° angular bins over a 3×3 partitioning. Figure 9 shows the confusion matrices of each method. Blank *et al.*

Table 4. The accuracies of the matching methods with respect to $N \times N$ grids (with 15° angular bins, no rectangle or torso elimination). We have compared 2×2 and 3×3 partition grids. Our results show that 3×3 grid is more effective when forming our oriented-rectangles based pose descriptor.

Matching Method	2×2	3×3
FrameVoting	0.9136	0.9630
GlobalHist	0.8765	0.9630
SVM	0.9012	0.9506
DTW	0.9136	**1.0000**

report classification error rates of 0.36% and 3.10% for this dataset. Recently, Niebles and Fei Fei [17] evaluate their hierarchial model of spatial and spatio-temporal features over this dataset, acquiring an accuracy of 72.8%.



(a) (b) (c) (d)

Fig. 9. Confusion matrices for each matching method using original rectangle distributions with no torso detection and 15° angular bins over a 3×3 grid. (a) Frame by frame voting : 1 jump sequence classified as bend, 1 one-hand wave sequence classified as jump-in-place and 1 run sequence misclassified as walk. (b) Global histogramming : 1 one-hand wave sequence misclassified as jump-in-place, 1 jumpjack sequence misclassified as two-hands-wave and 1 run sequence misclassified as walk. (c) SVM classification : 1 jump sequence is classified as bend, 2 run sequences classified as walk, 1 run sequence misclassified as jump. (d) DTW classification achieves 100% accuracy.

We should also note that frame by frame voting and global histogramming with our pose descriptor produce surprisingly good results. This suggests that we can still achieve satisfactory classification rates even if we ignore the time domain and look at the frames separately, or as a whole.

5 Conclusions and Future Work

In this paper, we have approached to the problem of human action recognition from a bag-of-features perspective and proposed a new pose-descriptor based on the orientation of body parts. Our pose-descriptor is simple and effective; we extract the rectangular regions from a human silhouette and form spatial oriented histogram of these rectangles. We show that by effective classification of such histograms, robust human action recognition is possible. We demonstrate

the effectiveness of our method over the dataset of Blank *et al.* [1]. Our results are directly comparable/superior than the results presented over this dataset.

The matching methods we present in this study suggest that we may not need a perfect modeling of the dynamics of human actions in order to reach satisfactory results. The questions behind the success of frame by frame voting or global histogramming methods are: "Do we really need dynamics of an action to recognize it correctly?", or "Can we simply recognize the actions by looking at representative frames or signatures of them?". Our experiments show that human pose encapsulates many useful information for the action itself, therefore, one can start with a good pose estimator, before going into the details of dynamics.

Future work includes application of our pose-descriptor to more complex datasets and still images. We also plan to explore the view-invariance case, by means of orthographic projections of rectangular regions. In addition, we will explore finer scale angular bins with varying spatial formations.

References

1. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV, pp. 1395–1402 (2005)
2. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE T. Pattern Analysis and Machine Intelligence 23(3), 257–267 (2001)
3. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 994–999. IEEE Computer Society Press, Los Alamitos (1997)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. I, pp. 886–893. IEEE Computer Society Press, Los Alamitos (2005)
5. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV 2003, pp. 726–733 (2003)
6. Fei-Fei, L., Perona, P.: A bayesian heirarcical model for learning natural scene categories. In: IEEE Conf. on Computer Vision and Pattern Recognition, IEEE Computer Society Press, Los Alamitos (2005)
7. Forsyth, D., Fleck, M.: Body plans. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 678–683. IEEE Computer Society Press, Los Alamitos (1997)
8. Forsyth, D., Arikan, O., Ikemoto, L., O'Brien, J., Ramanan, D.: Computational studies of human motion i: Tracking and animation. Foundations and Trends in Computer Graphics and Vision 1(2/3) (2006)
9. Freeman, W., Roth, M.: Orientation histograms for hand gesture recognition. In: International Workshop on Automatic Face and Gesture Recognition (1995)
10. Hong, P., Turk, M., Huang, T.: Gesture modeling and recognition using finite state machines. In: Int. Conf. Automatic Face and Gesture Recognition, pp. 410–415 (2000)
11. Hongeng, S., Nevatia, R., Bremond, F.: Video-based event recognition: activity representation and probabilistic recognition methods. Computer Vision and Image Understanding 96(2), 129–162 (2004)
12. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE transactions on systems, man, and cybernetics c: applications and reviews 34(3) (2004)

13. Ikizler, N., Forsyth, D.: Searching video for complex activities with finite state models. In: IEEE Conf. on Computer Vision and Pattern Recognition (2007)
14. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. Int. J. Computer Vision 43(1), 29–44 (2001)
15. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 246–253 (2006)
16. Monay, F., Gatica-Perez, D.: Modeling semantic aspects for cross-media image retrieval. IEEE T. Pattern Analysis and Machine Intelligence (accepted for publication)
17. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: IEEE Conf. on Computer Vision and Pattern Recognition, IEEE Computer Society Press, Los Alamitos (2007)
18. Oliver, N., Garg, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. Computer Vision and Image Understanding 96(2), 163–180 (2004)
19. Pinhanez, C., Bobick, A.: Pnf propagation and the detection of actions described by temporal intervals. In: DARPA IU Workshop, pp. 227–234 (1997)
20. Pinhanez, C., Bobick, A.: Human action detection using pnf propagation of temporal constraints. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 898–904. IEEE Computer Society Press, Los Alamitos (1998)
21. Polana, R., Nelson, R.: Detecting activities. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2–7. IEEE Computer Society Press, Los Alamitos (1993)
22. Ramanan, D., Forsyth, D., Zisserman, A.: Strike a pose: Tracking people by finding stylized poses. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. I, pp. 271–278. IEEE Computer Society Press, Los Alamitos (2005)
23. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Computer Vision 40(2), 99–121 (2000)
24. Siskind, J.M.: Reconstructing force-dynamic models from video sequences. Artificial Intelligence 151, 91–154 (2003)
25. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering object categories in image collections. In: Int. Conf. on Computer Vision (2005)
26. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: Int. Conf. on Computer Vision, pp. 1808–1815 (2005)
27. Wilson, A., Bobick, A.: Parametric hidden markov models for gesture recognition. IEEE T. Pattern Analysis and Machine Intelligence 21(9), 884–900 (1999)
28. Yu-Gang Jiang, C.-W.N., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Int. Conf. Image Video Retrieval (2007)