Image and Vision Computing 27 (2009) 1515–1526

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Histogram of oriented rectangles: A new pose descriptor for human action recognition

Nazlı Ikizler *, Pınar Duygulu

Dept. of Computer Engineering, Bilkent University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 17 January 2008
Received in revised form 23 January 2009
Accepted 12 February 2009

Keywords:
Action recognition
Human motion understanding
Pose descriptor

ABSTRACT

Most of the approaches to human action recognition tend to form complex models which require lots of parameter estimation and computation time. In this study, we show that, human actions can be simply represented by pose without dealing with the complex representation of dynamics. Based on this idea, we propose a novel pose descriptor which we name as Histogram-of-Oriented-Rectangles (HOR) for representing and recognizing human actions in videos. We represent each human pose in an action sequence by oriented rectangular patches extracted over the human silhouette. We then form spatial oriented histograms to represent the distribution of these rectangular patches. We make use of several matching strategies to carry the information from the spatial domain described by the HOR descriptor to temporal domain. These are (i) nearest neighbor classification, 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 to rectangular patches, (iii) a classifier-based approach using Support Vector Machines, and (iv) adaptation of Dynamic Time Warping on the temporal representation of the HOR descriptor. For the cases when pose descriptor is not sufficiently strong alone, such as to differentiate actions "jogging" and "running", we also incorporate a simple velocity descriptor as a prior to the pose based classification step. We test our system with different configurations and experiment on two commonly used action datasets: the Weizmann dataset and the KTH dataset. Results show that our method is superior to other methods on Weizmann dataset with a perfect accuracy rate of 100%, and is comparable to the other methods on KTH dataset with a very high success rate close to 90%. These results prove that with a simple and compact representation, we can achieve robust recognition of human actions, compared to complex representations.

1. Introduction

Human action recognition 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, current solutions are very limited, and understanding what people are doing remains unresolved.

Human action recognition has been a widely studied topic (for extensive reviews see [1,2]), but the solutions to the problem that have been submitted to date are very premature and still specific to the dataset at hand.

There are three key elements that define an action:

- pose of the body (and parts),
- speed of body motion (and parts),
- relative ordering of the poses.

We can formulate action recognition as a mixture of these three elements. The relative importance of these elements is based on the nature of the actions that we aim to recognize. For example, if we want to differentiate an instance of a "bend" action from a "walk" action, the pose of the human figure gives sufficient information. However, if we want to discriminate between "jog" and "run" actions, the pose alone may not be enough, due to the similarity in the nature of these actions in the pose domain, and in such cases, the speed information needs to be incorporated. Similarly, for recognizing "stand up" and "sit down" actions, the relative ordering of the poses will be important, since these two actions include same poses in reverse temporal orders.

Various attempts in action recognition literature try to model some or all of these aspects. For instance, methods based on spatio-temporal templates mostly pay attention to the pose of the human body, whereas methods based on dynamical models focus on modeling the ordering of these poses in greater detail.

We argue that the human pose encapsulates many useful clues for recognizing the ongoing activity. Actions can mostly be represented by configurations of the body parts, before building complex models for understanding the dynamics.

* Corresponding author. Tel.: +903122902094; fax: +903122664047.
E-mail addresses: inazli@cs.bilkent.edu.tr (N. Ikizler), duygulu@cs.bilkent.edu.tr (P. Duygulu).

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.02.002

Using this idea, we base the foundation of our method on defining the pose of the human body to discriminate actions, and by introducing a new pose descriptor, we want to evaluate how far we can go only with a good description of the pose of the body. We also evaluate how our system benefits from adding the remaining action components whenever necessary. Unlike most of the methods that use complex modeling of body configurations, we follow the analogy of Forsyth and Fleck [3], 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: the 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 is similar to the bag-of-words approach, where the images are represented by a collection of regions, ignoring their spatial relationships. The bag-of-words approach – which is adapted from text retrieval literature – has shown to be successful for object and scene recognition [4,5] and for annotation and retrieval of large image and video collections [6,7]. 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. In our approach, we use rectangles as our visual words and achieve vector quantization by histogramming over their orientation angles. However, our approach is basically different from bag-of-words. First, we are using the distribution of rectangular regions as opposed to complex visual words. Second, we place a grid over these rectangles to capture their spatial layout.

In this study, our main contribution is to show how a good pose descriptor can boost the performance of action recognition. We introduce a novel pose descriptor which is based on candidate rectangular regions over the human body. We show that using our pose descriptor, we can recognize human actions even in complicated settings.

In the rest of the paper, we first cover the literature on human action recognition within a brief overview. Then, we present the details of our pose descriptor, which represents the human figure as a distribution of oriented rectangular patches. After that, we list the matching methods that can be applied to our pose descriptor for efficient identification of human actions. These are, namely, nearest neighbor classification, global histogramming, SVM classification and Dynamic Time Warping. We test our system with different configurations and compare the results to state-of-art action recognition methods. We also provide the run time evaluations of our system. After reporting comprehensive experiments and their results, we conclude our discussion with future research directions.

2. Related work

There are three major approaches to human action understanding in videos. The first one is to use temporal logics to represent crucial order relations between states that constrain activities. Examples of such approaches include Pinhanez and Bobick [8,9] who described a method based on interval algebra. In addition, Siskind [10] described methods to infer activities related to objects using a form of logical inference.

The second general approach to recognizing human motion is to use models of dynamics. Such models can be constructed as hidden markov models [11–13], conditional random fields [14], or finite state models [15,16]. These models rely on describing the details of the action dynamics. Although these methods capture the details for action dynamics in a more natural way, the foremost shortcoming of such methods is the need for a great deal of training data to build effective and reliable models. Ikizler and Forsyth [17] show how to make use of motion capture data in such a case.

Third main approach is to use spatio-temporal templates to identify instances of activities. Spatio-temporal patterns date back to Polana and Nelson [18]. Later on, thinking actions as such spatio-temporal templates were made famous by Bobick and Davis [19]. They introduced Motion-Energy-Image and Motion-History-Image templates for recognizing different motions. Efros et al. [20] use a motion descriptor based on optical flow of a spatio-temporal volume. Blank et al. [21] also define actions as space–time shapes, making use of Poisson distributions to define the details of such shapes. A very recent approach based on a hierarchical use of spatio-temporal templates tries to model the ventral system of the brain to identify actions [22].

Recently, the "bag-of-words" approaches originated from text retrieval research are being adapted to action recognition. These studies are mostly based on the idea of forming codebooks of "spatio-temporal" features. Laptev and Lindeberg first introduced the notion of "space–time interest points" [23] and used SVMs to recognize actions [24]. Dollár et al. extracted cuboids via separable linear filters and formed histograms of these cuboids to perform action recognition [25]. Niebles et al. applied a pLSA approach over these patches to perform unsupervised action recognition [26]. Recently, Wong et al. proposed using pLSA method with an implicit shape model to infer actions from spatio-temporal codebooks [27]. More recent work on human action recognition includes recognizing the "actions in the wild" – actions with more challenging settings like in the case of movies [28,29]. Using self similarities for action description is another approach which has been introduced recently [30].

Histogramming is an old trick that has been frequently used in computer vision research. For action recognition, Freeman and Roth [31] used orientation histograms for hand gesture recognition. Recently, Dalal and Triggs used histograms of oriented gradients (HOGs) for human detection in images [32], which is shown to be quite successful.

Our approach falls into the category of action recognition via spatial-temporal templates. We do not use an explicit generative model for action dynamics. Instead, we make use of the discriminative classification approaches wrapped around our novel pose descriptor.

An earlier version of this paper appeared in Human Motion Workshop held in conjunction with ICCV 2007 [33]. In [33], we introduced the basic pose descriptor and presented preliminary results on the Weizmann dataset. In this manuscript, we introduce an extended version of our pose descriptor (HOR), which is histogram of oriented rectangles over a window (HORW), and we extend the classification schemes that can be used with these pose descriptors. We describe a two-level classification procedure, where the speed of the person is used as an early rejection step for SVM classification. In Section 5, we present comprehensive results on the Weizmann and the KTH datasets, and elaborate on certain parameter choices. We also compare our method to its natural counterpart, i.e. Histograms of Oriented Gradients [32] and provide the runtime evaluations of our method.

3. Histogram of oriented rectangles as a new pose descriptor

Following the body plan analogy of Forsyth and Fleck [3], we represent the human body as a collection of rectangular patches and we base our motion understanding approach on the fact that

the orientations and positions of these rectangles change over time with respect to the 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, a tracker identifies the location of the subject. Then, the bounding box around its silhouette is extracted. This bounding box is then divided into $N \times N$ 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 based on the orientations of the rectangles in each spatial bin. This process is depicted in Fig. 1.

In the ideal case, single rectangles that fit perfectly to the limb areas should give enough information about the pose of the body. However, finding those perfect rectangles is not straightforward and is very prone to noise. Therefore, in order to eliminate the effect of noise, we use the distribution of candidate rectangular regions as our feature. This gives a more precise information about the most probable locations of the fittest rectangles.

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.

3.1. Extraction of rectangular patches

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 [2] for a detailed overview of the subject). In our experiments, we use a background subtraction scheme 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. [34]. The search is performed using 12 tilting angles, which are 15° apart, covering a search space of 180°. Note that since we do not have the directional information of these rectangle patches, orientations do not cover 360°, but its half. To tolerate the differences in the limb sizes and in the varying camera distances to the subject, we perform the rectangle convolution over multiple scales.

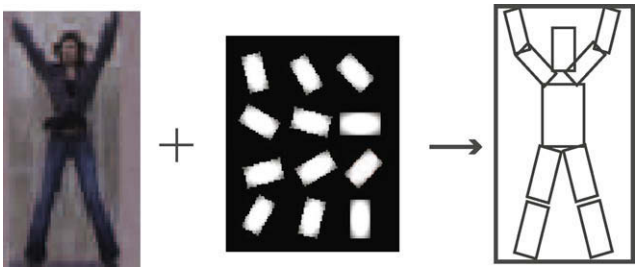


Fig. 2. The rectangular filtering process is shown. We use zero-padded Gaussian filters with 15° tilted orientations over the human silhouette. We search over various scales, without discriminating between different body parts. The perfect rectangular search for the given human subject would result in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

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}:

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

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

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

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. We then select the k of the remaining candidate regions of each scale by random sampling (we used $k = 300$).

3.2. Describing pose as 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 compute 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 an $N \times N$ 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 (we assume a fixed body and leg height, where each silhouette is normalized to the fixed body height). The area covering the silhouette is divided into equal-sized bins from bottom to top and left to right (see Fig. 3 for details). Note that, in this way, we give some space to the top part of the head, to allow action space for the arms (for actions like reaching, waving, etc.).

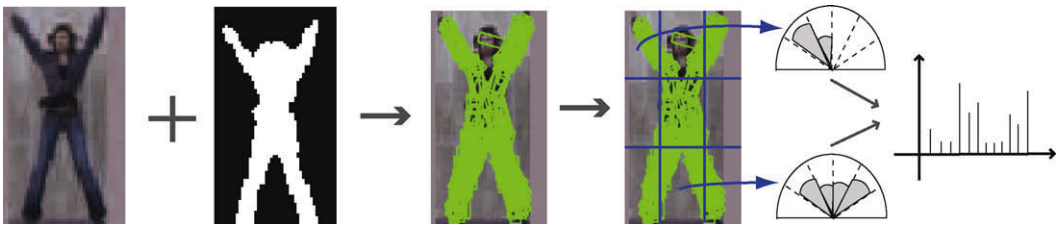


Fig. 1. Here, the feature extraction stage of our approach is shown (this figure is best viewed in color). First, the human figure in each frame is extracted using background subtraction or an appropriate tracker. Using these silhouettes, we search for the rectangular patches that can be candidates for the limbs. We do not discriminate between legs and arms. 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 from each subregion.

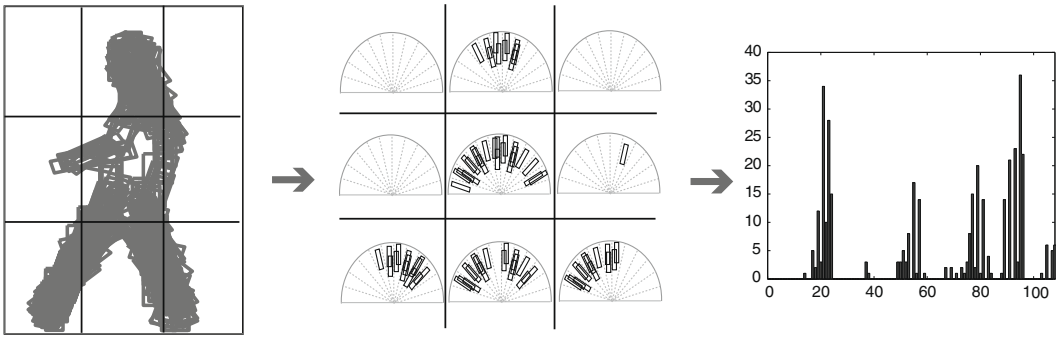


Fig. 3. Details of histogram of oriented rectangles (HORs). The bounding box around the human figure is divided into an $N \times N$ grid (in this case, 3×3) and the HORs from each spatial bin are shown. The resulting feature vector is a concatenation of the HORs from each spatial bin.

We also evaluate the effects of using different granularities in orientation bins and sparser grids over the human body, which have more concise feature representations, but coarser detail of the human pose. We show the corresponding results in Section 5.

3.3. Capturing local dynamics

In action recognition, there may be times where one cannot discriminate two actions by just looking at single poses. In such cases, an action descriptor based purely on shape is not enough and temporal dynamics must be explored. To incorporate temporal features, HORs can be calculated over snippets of frames rather than single frames. More formally, we define histograms oriented rectangles over a window of frames (HORW), such that the histogram of the ith frame will be

$$HORW(i) = \sum_{k=i-n}^{i} HOR(k), \tag{2}$$

where n is the size of the window.

By using HORs over a window of frames like this, we capture local dynamics information. In our experiments, we observe that, using HORW is more useful especially to discriminate actions like "jogging" and "running", which are very similar in pose domain, but different in speed. Therefore, over a fixed length window, the compactness of these two actions will be different. We evaluate the effect of using HOR vs HORW in greater detail in Section 5.

4. Recognizing actions with histograms of oriented rectangles

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

4.1. Nearest neighbor classification

The simplest scheme we utilize is to perform matching based on single frames (or snippets of frames in the case of HORWs), ignoring the dynamics of the sequence. That is, for each test instance frame, we find the closest frame in the training set and assign its label as the label of the test frame. We then employ a voting scheme throughout the whole sequence. This process is shown in Fig. 4.

The distance between frames is computed using χ^2 distance between the histograms (as in [35]). Each frame with the histogram H_i is labeled 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 bounding box center of the human silhouette may result in a different binning of the rect-

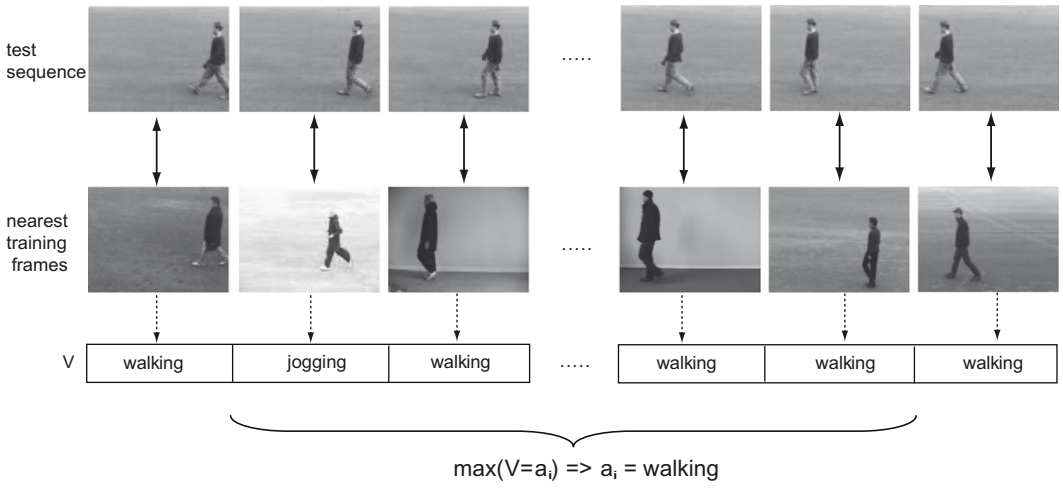


Fig. 4. Nearest neighbor classification process for a walking sequence. The pose descriptor of each frame is compared to that of the training set frames and the closest frame's class is assigned as a label to that frame. The resulting is a vote vector, where each frame contributes with a vote and the majority class of the votes is the recognized action label for that sequence.

angles and, therefore, large fluctuations in distance. One can utilize Earth Mover's Distance [36] or Diffusion Distance [37], which are shown to be more efficient for histogram comparison in the presence of such shifts, by taking the distances between bins into account at the expense of higher computation time.

4.2. Global histogramming

Global histogramming is similar to the Motion Energy Image (MEI) method proposed by Bobick and Davis [19]. In this method, we sum up all spatial histograms of oriented rectangles through the sequence, and form a single compact representation for the entire video. This is simply done by collapsing all time information into a single dimension by summing the histograms and forming a global histogram H_{global} such that

$$H_{global}(d) = \sum_{t} H_t(d) \tag{4}$$

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 label of the closest match is reported. The corresponding global images are shown in Fig. 5. These images show that for each action (of Weizmann dataset in this case), even a simple representation like global histogramming can provide useful interpretations.

Global histogramming is most effective when the sequences are long enough to capture the related pose information with each action. When the sequences are long enough, i.e. there are multiple cycles for each action, the accumulated histogram approximates the overall pattern of the actions more accurately. However, if the sequences are short and the number of cycles are different, the accumulated histogram is likely to approximate the structure of the action poorly.

4.3. 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. This process is depicted in Fig. 6. 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

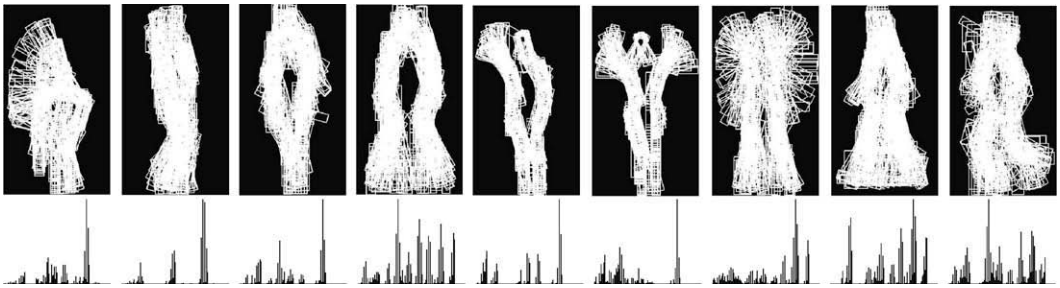


Fig. 5. Global histograms are generated by summing up all the sequence and forming the spatial histograms of oriented rectangles from these global images. In this figure, global images after the extraction of the rectangular patches are shown for nine separate action classes. These are `bend`, `jump`, `jump in place`, `gallop sideways`, `one-hand wave`, `two-hands wave`, `jumpjack`, `walk` and `run` actions. These images resemble the Motion Energy Images introduced by [19], however we do not use these shapes. Instead, we form the global spatial histogram of the oriented rectangles as our feature vector. These global spatial histograms are shown below each motion image, respectively.

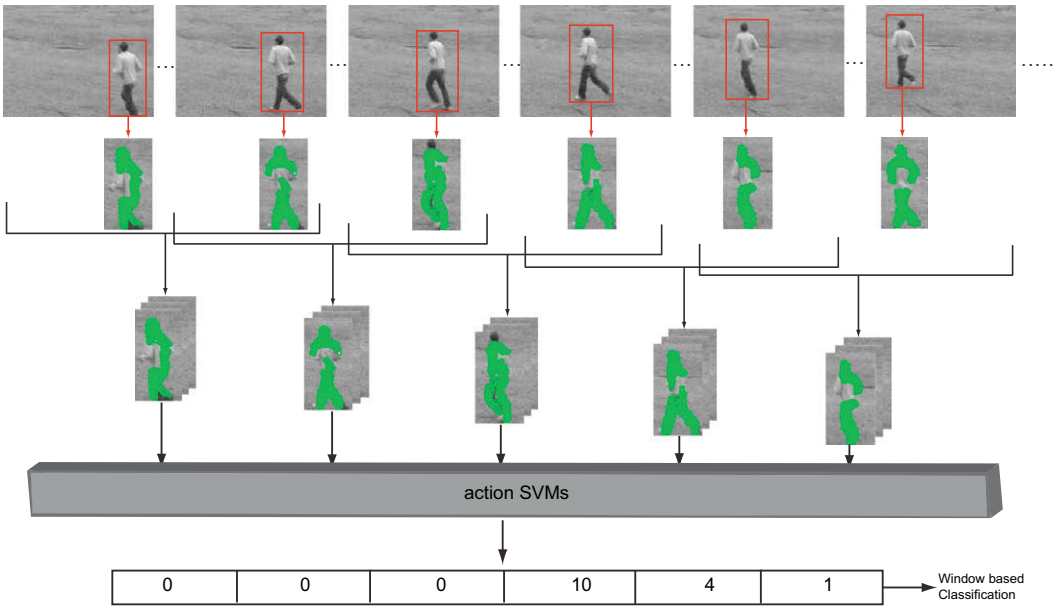


Fig. 6. SVM classification process over a window of frames.

results with $k = 15$, and $o = 3$). The whole sequence is then labeled with the most frequent action class among its chunks.

4.4. Dynamic time warping

Since the temporal durations of the actions are not uniform, comparing sequences is not straightforward. In the case of human actions, the same action can be performed at different speeds, causing the sequence to be expanded or shrunk 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 (for more details, see [38]). The time axes are warped in such a way that samples of the corresponding points are aligned.

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

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

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 Eq. 3.

We use dynamic time warping along each dimension of the histograms separately. As shown in Fig. 7, we take each 1D series of the histogram bins of the test video X and compute the DTW distance $D(X(d), Y(d))$ to the corresponding 1D 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.

4.5. Classification with speed information

When shape information is not enough, we can also use speed information as a prior for action classes. Suppose we want to discriminate two actions: "handwaving" versus "running". If the velocity of the person in motion is equal to zero, the probability that he has been running is quite low.

Based on this observation, we propose a two-level classification system. In the first level, we calculate mean velocities of the training sequences and fit a gaussian to each action in action set $A = \{a_1 \cdots a_n\}$. Later on, given a test instance, we compute the posterior probability of each action $a_i \in A$ over these gaussians, and if the posterior probability of a_i is greater than a threshold t (we use a loose bound $t = 0.1$), then we add a_i to the probable set S of actions for that sequence. After this preprocessing step, as the second level, we evaluate only the outputs of the SVMs for actions $a_k \in S$, and we take the maximum response from this subset of SVM classifiers as our classification decision. This process is shown in Fig. 8.

5. Experimental results

5.1. Datasets

We test the effectiveness of our method over two state-of-the-art datasets. The first is the Weizmann dataset and second is the KTH dataset, which are the current benchmark datasets in action recognition literature.

Weizmann dataset: This is the dataset that Blank et al. introduced in [21]. We used the same set of actions as in [21], which is a set of nine actions: walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place and jumping jack. Example frames from this dataset are shown in Fig. 9. We used the extracted masks provided to localize the human figures in each image. These masks were obtained using background subtraction. We test the effectiveness of our method using leave-one-out cross validation.

KTH dataset: This dataset has been introduced by Schuldt et al. in [24]. It is more challenging than the Weizmann, covering 25 subjects and four different recording conditions of the videos. There are six actions in this dataset: boxing, handclapping, handwaving, jogging, running and walking. One additional challenge of this dataset comes from the set of actions available; there are two very similar actions – jogging and running – in this dataset. Example frames from the KTH dataset are shown in Fig. 10.

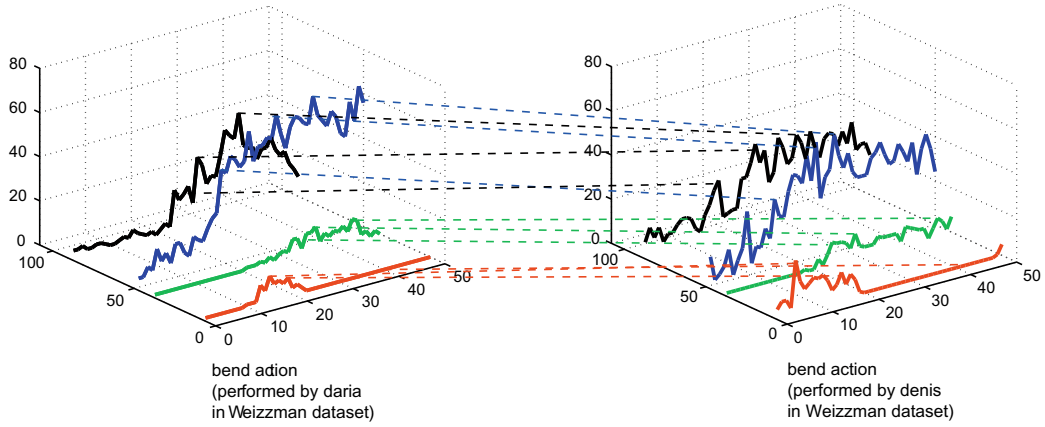


Fig. 7. Dynamic Time Warping (DTW) over 2D histograms: we compute DTW distances between the histograms by evaluating the DTW cost over single dimensions separately and summing up all costs to get a global distance between sequences. Here, histograms of two bend actions performed by different actors are shown. We try to align these sequences along each histogram dimension by DTW and report the 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.

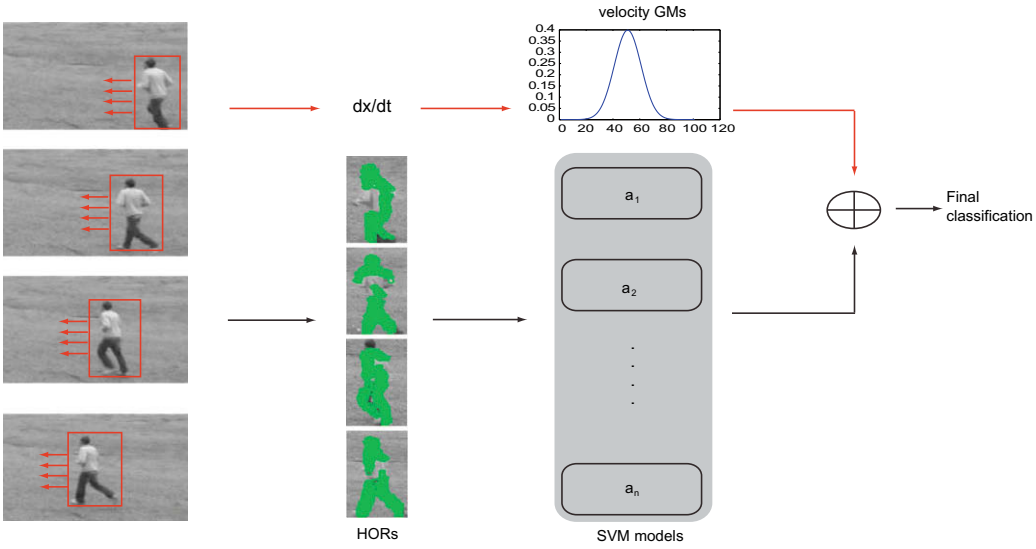


Fig. 8. Two-level classification of actions based on mean horizontal velocity and histograms of oriented rectangles. First, the velocity of the subject is calculated throughout the entire video. We evaluate the posterior probability of this velocity and determine the probable set of actions for that video. Then, based on this probable set of actions, we look at the responses from corresponding SVM classifiers and take the maximum response as the classification label of that video.

Fig. 9. Example frames from the Weizmann dataset introduced in [21].

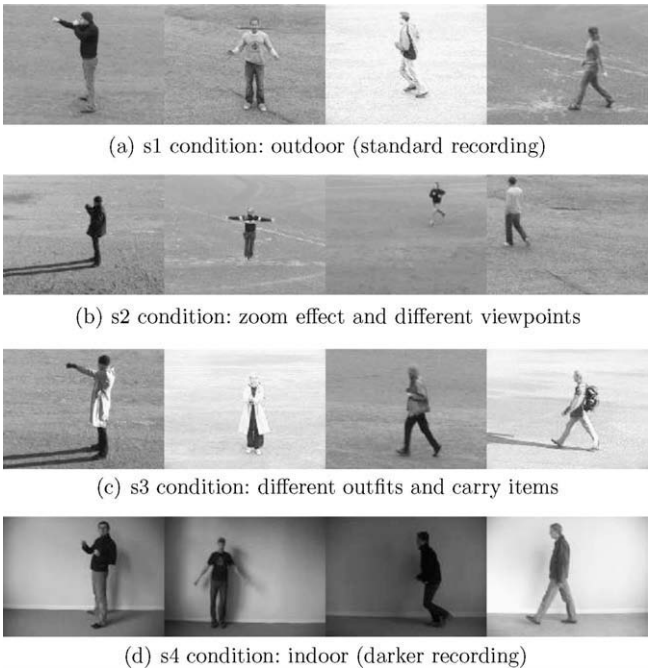


Fig. 10. Example frames from the KTH dataset introduced in [24].

Since the recording conditions of the videos in the KTH dataset are not stable, and there is considerable amount of camera movement in some cases, silhouette extraction in this dataset is not straightforward. For this reason, we follow an ad hoc approach and make use of several cues like gradient responses, for a good extraction of the foreground human figure. In the KTH dataset, despite the camera movement and zoom effect, the backgrounds of the sequences are relatively simple. We used this fact to localize the human figure, and then applied background subtraction to the localized image region.

To be more specific, we use the observation that the width of the person inside the image is mostly specified by the horizontal edges (like a forward arm). Similarly, the height of the person is specified by the vertical edges. For determining the width, we first compute the gradient responses F_y of each image and take the projections by marginalizing over the y dimension. This gives us the density of the horizontal edges along the x dimension. Similarly, for finding the height of the person, we marginalize the gradient responses of vertical edges (F_x) over the x dimension, to find the density of vertical edges at each y coordinate. Then, we threshold these marginal densities as follows:

$$\sum_i F_x(i,j) > \tau_x, \tag{7}$$

$$\sum_j F_y(i,j) > \tau_y, \tag{8}$$

where i is the horizontal index and j is the vertical index, and τ_x and τ_y are the threshold values empirically chosen to be high enough to get rid of the effect of the noise. We then fit a bounding box by taking the minimum and maximum indices that satisfy the above conditions. We apply background subtraction to the region specified by the bounding box coordinates to get the final silhouettes.

The resulting silhouettes are not perfect, but realistic. Some of the example silhouettes are shown in Fig. 11. The successful results on these noisy silhouettes prove that our method does not heavily depend on perfect silhouettes. We should note that, better silhouettes will give higher accuracy rates eventually.

5.2. Best configuration of the pose descriptor

We first evaluate the performance of the pose descriptor with respect to its configuration. There are several choices that can be

made while forming the HOR pose descriptor. These are (a) granularity of the angular bins, i.e. number of orientations for the rectangle detection, (b) number of spatial bins, (c) form of the histogram and (d) choice of torso exclusion.

In the following, we first exploit the effect of using different configurations over the Weizmann dataset. Based on the empirical evaluation, we determine the best configuration for the Weizmann dataset and continue the rest of the experiments using the same configuration.

5.2.1. Granularity of angular bins

We first evaluated the choice of orientation angles when forming the histogram. Table 1 shows the results using different angular bins. Not surprisingly, we see that there is a slight loss of information when we go from fine level orientations (i.e. 15° bins) to a coarser level (30°). More interestingly, if we do not use angular binning, i.e. ignore the orientations, and just utilize the histogram of rectangles falling into each spatial grid, we may still capture a valuable amount of information (180° case). This confirms that describing the human figure as a collection of rectangles is a sensible approach, and even the spatial distribution of the rectangular regions over the silhouette provide quite rich information about the available action. If we look at the orientation of these rectangles besides the spatial distributions, we acquire more detail and higher accuracy about the action of the body.

5.2.2. Grid size

When forming the histograms of oriented rectangles, we place an $N \times N$ grid over the silhouette of the subject and form orientation histograms for each grid region. The choice of N affects the size of the feature vector (thus execution time of the matching), and the level of detail of the descriptor. Table 2 compares the use of different spatial grids. The 1×1 grid implies that we do not use any spatial binning and we take the silhouette as a whole. Not surprisingly, ignoring the spatial layout and binning only over orientations is not satisfactory, since spatial layout of the rectangles provides useful cues for discriminating different parts of the body.

Table 1
The accuracies of the matching methods with respect to angular bins (over a grid of 3×3). 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. 180° HORs ignore the orientation information of the rectangles and performs binning based on the spatial distribution of the rectangles over the silhouette. Surprisingly, even the spatial distribution of the rectangular regions provide quite rich information about the available action.

	5° (%)	10° (%)	15° (%)	30° (%)	180° (%)
NearestNeighbor	90.12	93.83	96.30	95.06	92.59
GlobalHist	60.49	80.25	96.30	93.83	85.19
SVM	95.06	93.83	97.53	93.83	92.59
DTW	85.19	93.83	100	95.06	91.36

Table 2
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 the 3×3 grid is more effective when forming our oriented-rectangles based pose descriptor.

	1×1 (%)	2×2 (%)	3×3 (%)	4×4 (%)
NearestNeighbor	64.20	91.36	96.30	93.83
GlobalHist	55.56	87.65	96.30	82.72
SVM	80.25	90.12	97.53	90.12
DTW	70.37	91.36	100	91.36

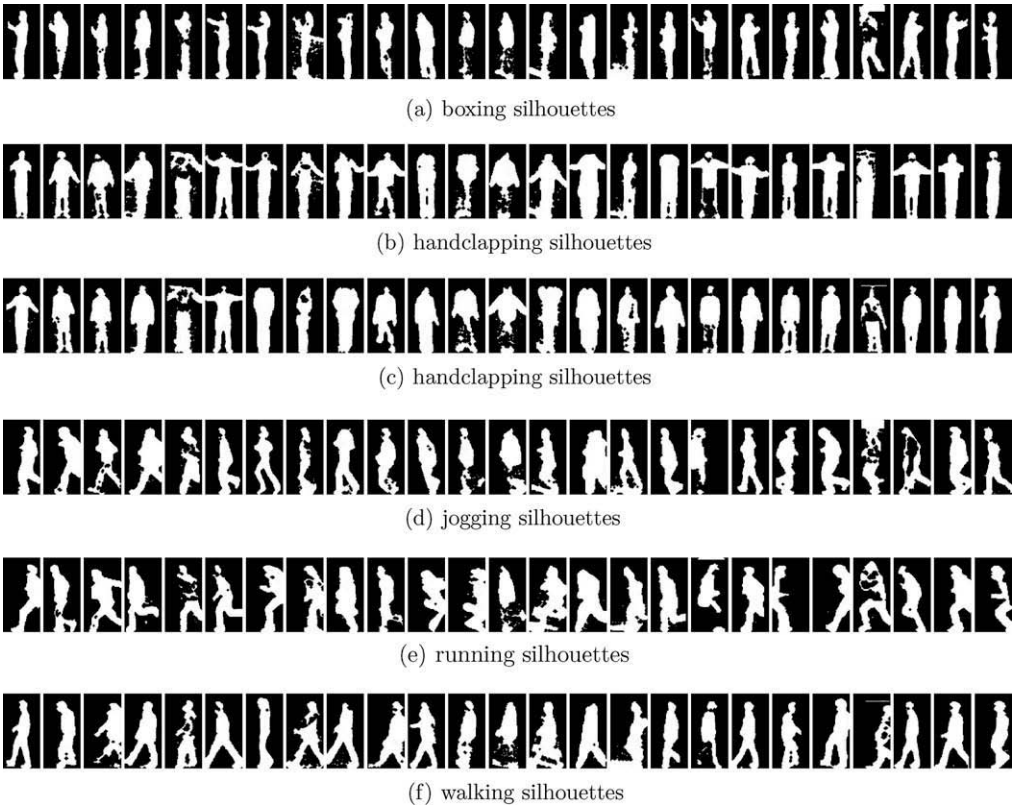


Fig. 11. Extracted silhouettes from the KTH dataset in s1 recording condition. Each single silhouette image corresponds to a different actor in the dataset (a total of 25 actors). The silhouettes are quite imperfect, due to the difficulty in foreground/background discrimination in this dataset.

Our results over this dataset indicate that the 3×3 grid gives the best performance. However, if execution time is crucial, choice of $N = 2$ (2×2 grid) will still work to a certain degree of performance. For $N = 4$, the classification accuracy is lower due to the loss in generalization power, and the classification time increases due to the increase in feature dimensionality. One can try further levels of partitioning, even form pyramids of these partitions. However, overly dense partitioning will not make sense, since the subregions have to be large enough to contain a reasonable amount of rectangular patches.

We also evaluated the grid size selection and the granularity of angular bins together using global histogramming method. Fig. 12 shows the corresponding results. As it can also be seen here, the best configuration for these parameters are using 3×3 grid with 15° angular bins.

5.2.3. Torso detection

One can perform a special search for the torso rectangle, which is considerably larger than limb rectangles, omit this torso region while searching for the remaining body parts and then form rectangular histograms. An example case for this kind of rectangular search is given in Fig. 13. Here, by applying a larger rectangular filter and computing the mean among the responses, we localize the torso region. Then, we exclude this region and apply rectangle filtering in order to find candidate limb areas and base our pose descriptor on these areas only.

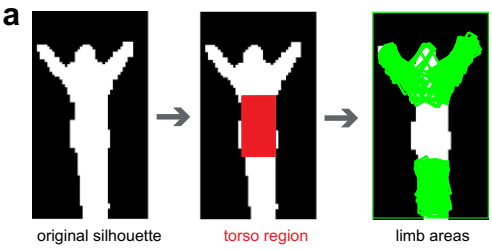
In Fig. 13, we show the effect of torso detection on the overall accuracies. We observe that with global histogramming methods, torso detection and exclusion helps; however, SVM and DTW classifiers suffer from slight performance degradation. So, we conclude that explicit torso detection is not necessary and extracting the HOR descriptors from the whole silhouettes is more informative.

5.3. Classification results and discussions

After finding the best configuration of the pose descriptor over the Weizmann dataset, we evaluate the effect of using different classification techniques. For this, we use the same configuration found in the previous section which is a 3×3 grid over 15° angular bins as our HOR configuration. These parameters depend on the proportions of the human body and are expected to scale well in fullbody action datasets where the videos have the similar level of detail.

The overall results over two datasets are shown in Table 3. For the Weizmann dataset, where actions are mostly differentiable based on their shape information, applying DTW over HOR descriptors gives the best results. However, on the more complex KTH dataset, we need to make use of the velocity information, because shape is mostly not enough, especially in the presence of noise introduced by imperfect silhouettes. In the KTH dataset, best results are achieved by using two-level classification with SVM models (v+SVM) using HORWs as features.

In Figs. 14 and 15, we present the confusion matrices of our method in Weizmann and KTH datasets, respectively. On the Weiz-



b

	No Torso	Torso
NN	96.30%	96.30%
GlobalHist	96.30%	91.36%
SVM	97.53%	95.06%
DTW	100%	98.77%

Fig. 13. Rectangle detection with torso exclusion (best viewed in color). In (a), 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 analyzed for candidate limb areas. In (b), the accuracies of the matching methods with respect to torso exclusion are given (using 15° angular bins and 3×3 grid). We can say that torso detection degrades the performance, so using the whole silhouette for candidate rectangle regions results in higher performance.

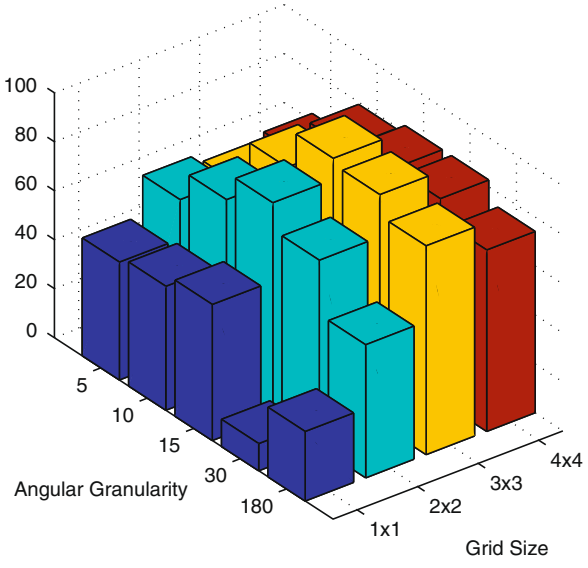


Fig. 12. The evaluation of grid size and the granularity of angular bins together. Here, global histogramming method is used for histogram comparison and the best performance is 96.30% achieved using 3×3 grid with 15° angular bins.

Table 3
Overall performance of the matching methods over the Weizmann and KTH datasets. The best results achieved are shown in bold.

	Weizmann (%)	KTH (%)
NearestNeighbor		
HOR	96.30	75.46
HORW	97.53	72.22
GlobalHist		
HOR	96.30	71.76
HORW	69.14	57.41
SVM		
HOR	97.53	77.31
HORW	95.06	85.65
DTW		
HOR	**100**	74.54
HORW	96.30	78.24
v+SVM		
HOR	98.77	81.48
HORW	95.06	**89.35**
v+DTW		
HOR	**100**	81.02
HORW	98.77	83.8

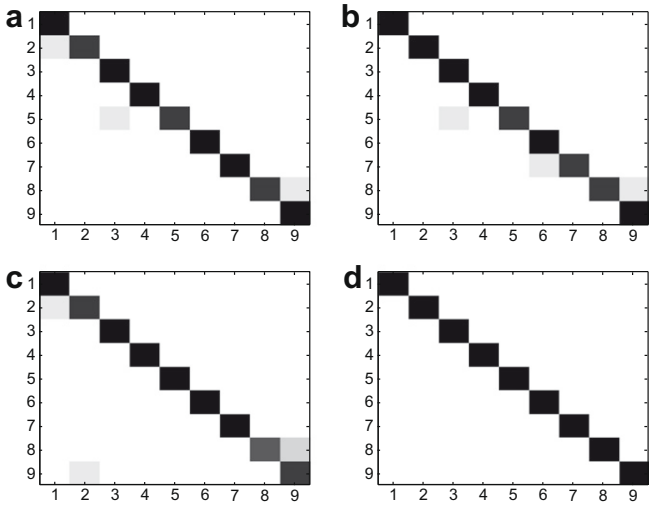


Fig. 14. Confusion matrices for each matching method over the Weizmann dataset using original rectangle distributions with no torso detection and 15° angular bins over a 3×3 grid. (a) *Nearest neighbor voting:* one jump sequence classified as bend, one one-hand wave sequence classified as jump-in-place and one run sequence misclassified as walk. (b) *Global histogramming:* one one-hand wave sequence misclassified as jump-in-place, one jumpjack sequence misclassified as two-hands wave and one run sequence misclassified as walk. (c) *SVM classification:* one jump sequence is classified as bend, two run sequences classified as walk, one run sequence misclassified as jump. (d) DTW classification achieves 100% accuracy.

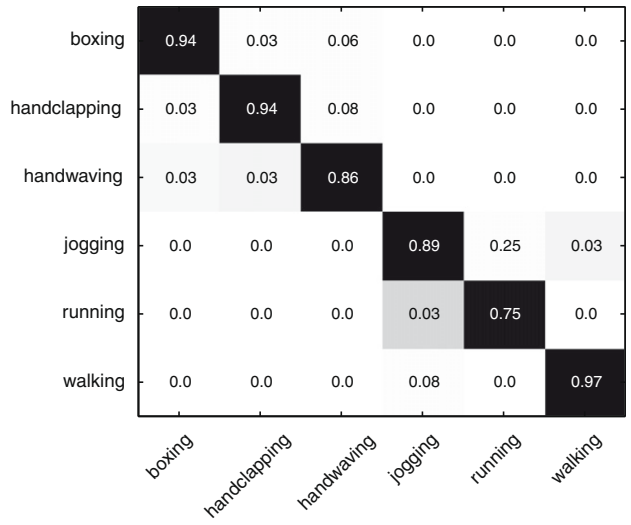


Fig. 15. Confusion matrix for classification results of the KTH dataset. Most of the confusion occurs between run and jog actions, which is quite comprehensible.

mann dataset, 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.

However, in the KTH dataset, a simple alignment using DTW is not sufficient, because in DTW we lose valuable time information by warping the time axes, and only pay attention to the ordering of the poses. The KTH dataset introduces additional challenge by including very similar actions like jogging and running, which need speed features to achieve better separation. Therefore, in this dataset, v+SVM classification performs best.

We should also note that, especially on the Weizmann dataset, nearest neighbor voting per frame 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 to a certain degree and look at the frames separately, or as a whole.

5.4. Comparison to other methods and HOGs

We reach a perfect accuracy (100%) over the Weizmann action dataset, using 15° angular bins over a 3×3 spatial partitioning with DTW or v+DTW methods. We present comparison of our results over this dataset in Table 4. Blank et al. report classification error rates of 0.36% and 3.10% for this dataset. Recently, Niebles and Fei-Fei [39] evaluate their hierarchial model of spatial and spatio-temporal features over this dataset, acquiring an accuracy of 72.8% and Wang and Suter [40] used FCRF over a grid-based feature, resulting in an accuracy of 97.78%.

In Table 5, we compare our descriptor's performance to current results on the KTH dataset. We should note that the numbers in Table 5 are not directly comparable, because the testing settings are different. Some of the approaches use leave-one-out cross-validation, whereas some others use different splitting of train and test data. We use the train and test sets provided in the original release of the dataset by Schuldt et al. [24]. Overall, we can say that our approach is among the top-ranking approaches in the literature regarding this dataset.

We also compare our approach to the HOGs, which is also based on histogramming and therefore is a natural counterpart to our approach. The HOG method has been recently proposed by Dalal and Triggs [32]. They have used gradient orientations to detect humans in still images, and their approach has been shown to be quite successful.

We used provided HOG implementation in order to extract the HOGs in the KTH dataset. While doing this, we omit the human detection phrase and we compute HOG features directly over the bounding box of the extracted silhouettes, using parameters *cell size* = 8 and *#of cells* = 2. This gives a feature vector of size 288, which is computationally very expensive, especially when used with SVMs over window of frames. In order to cope with this, we reduce the size of the HOG vectors by applying PCA and using the projections over the 80 principal components.

Table 6 shows the comparison results of HOGs and HORs using three of the most successful matching methods over the KTH dataset. As the results indicate, using HORs as opposed to HOGs gives a better classification performance in all matching techniques.

5.5. Choosing the window size

In our experiments, we empirically decide on the window size for HORW calculation. Fig. 16 shows the effect of selecting different window sizes for KTH dataset. On this dataset, computing HORW features using $\geqslant 9$ consecutive frames gives the best results. The classification accuracy rises as we increase the window size upto nine frames, whereas it remains unaffected by the increase from

Table 4
Comparison of our method to other methods that have reported results over the Weizmann dataset.

Method	Accuracy (%)
Our method	**100**
Blank et al. [21]	99.64
Jhuang et al. [22]	98.8
Wang et al. [40]	97.78
Niebles et al. [39]	72.8

Table 5
Comparison of our method to other methods that have reported results over KTH dataset.

Method	Accuracy (%)
Jhuang et al. [22]	91.7
Wong et al. [27]	91.6
Our method	**89.4**
Niebles et al. [26]	81.5
Dollár et al. [25]	81.2
Ke et al. [41]	80.9
Schuldt et al. [24]	71.7

Table 6
Comparison to HOG feature based action classification over the KTH dataset. We extract the HOGs using available implementation and optimal parameters (cellsize = 16, numcel = 1, descstride = 16). To be more computationally efficient, we applied PCA over HOGs and took the projection over 80 strongest principle vectors.

	HOG	HOGW	HOR	HORW
SVM %	76.85	77.78	77.31	85.65
DTW %	67.59	81.94	74.54	78.24
v+SVM %	82.41	86.11	81.48	89.35

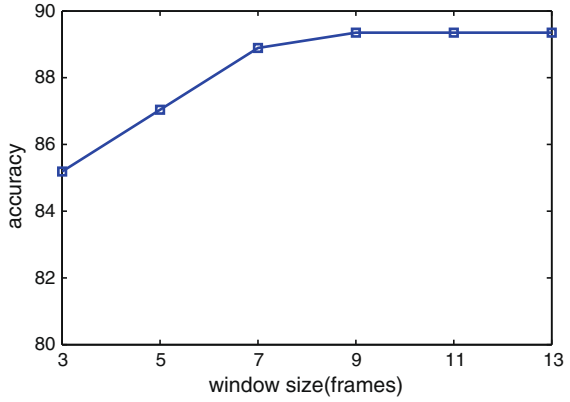


Fig. 16. The evaluation of choosing different window sizes for HORW calculation on KTH dataset. The results presented here are for v+SVM method. The best performance is achieved with window size ≈9 in this dataset.

that point on. Using HORW features provides more robustness to the noise in silhouettes, therefore, we can say that, in the KTH dataset, enough information about the actions are collected when using nine frame windows. This optimal value may vary for different datasets and on less noisy data, the best performance may be achieved in smaller window sizes.

5.6. Computational evaluation

The run-time evaluation of our approach is 2-fold. First is the phase of rectangle extraction. Rectangle extraction consumes around 1 second per frame.

The second phase is the matching part. The computational evaluation of the methods (implemented in MATLAB R2007a without code optimization) is presented in Table 7. These are the per-frame running times (in milliseconds) of corresponding methods over the Weizmann dataset. DTW is the most time-consuming method among others, whereas global histogramming takes the least amount of time. SVM classification has very managable time constraints and is preferable if the running time is an important consideration of the system.

Table 7
Run time evaluations for different matching techniques using HORs. The results presented here are the per-frame running times over the Weizmann dataset.

	NearestNeighbor	GlobalHist	SVM	DTW	v+SVM	v+DTW
msec	70.58	3.82	32.0	81.84	35.49	82.47

6. Conclusions and future work

In this paper, we have approached the problem of human action recognition 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 a spatial oriented histogram of these rectangles. We show that, by effective classification of such histograms, reliable human action recognition is possible. We demonstrate the effectiveness of our method over the state-of-the-art datasets in action recognition literature, which are the Weizmann and the KTH datasets. Our results are directly comparable and even superior to the results presented over these datasets.

Our experiments show that the human pose encapsulates many useful pieces of information about the action itself, and therefore one can start with a good pose estimator, before going into the details of dynamics. When pose itself is not enough for discriminating between actions, we show how to boost the performance by including simple velocity features and build a hierarchical model on top of our classification scheme. We show how we can obtain efficient action recognition with the minimum of dynamics information and complex modelling. Result is an intuitive and fast action recognition system with high accuracy rates, even in challenging conditions.

One shortcoming of our approach is its dependence over silhouette extraction. We observed that most of the confusion, especially in the KTH dataset, occurs because of the imperfect silhouettes. However, we should also note that, even with imperfect silhouettes, our method achieves high recognition rates which shows our method's robustness to noise. We argue that better silhouettes will result in higher accuracy rates eventually.

Future work includes application of our pose descriptor to still images. We also plan to extend our pose descriptor to cover the view-invariance case, by means of orthographic projections of rectangular regions.

Acknowledgements

This work has been supported by TUBITAK grants 104E065, 104E077 and 105E065.

References

[1] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews 34 (3) (2004).
[2] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, D. Ramanan, Computational studies of human motion I: tracking and animation, Foundations and Trends in Computer Graphics and Vision 1 (2/3) (2006).
[3] D. Forsyth, M. Fleck, Body plans, in: IEEE Conf. on Computer Vision and Pattern Recognition, 1997, pp. 678–683.
[4] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2005.
[5] J. Sivic, B. Russell, A. Efros, A. Zisserman, W. Freeman, Discovering object categories in image collections, in: Int. Conf. on Computer Vision, 2005.
[6] F. Monay, D. Gatica-Perez, Modeling semantic aspects for cross-media image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1802–1817.
[7] C.W.N. Yu-Gang Jiang, J, Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Int. Conf. Image Video Retrieval, 2007.
[8] C. Pinhanez, A. Bobick, Pnf propagation and the detection of actions described by temporal intervals, in: DARPA IU Workshop, 1997, pp. 227–234.
[9] C. Pinhanez, A. Bobick, Human action detection using pnf propagation of temporal constraints, in: IEEE Conf. on Computer Vision and Pattern Recognition, 1998, pp. 898–904.

[10] J.M. Siskind, Reconstructing force-dynamic models from video sequences, Artificial Intelligence 151 (2003) 91–154.
[11] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, in: IEEE Conf. on Computer Vision and Pattern Recognition, 1997, pp. 994–999.
[12] A. Wilson, A. Bobick, Parametric hidden Markov models for gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (9) (1999) 884–900.
[13] N. Oliver, A. Garg, E. Horvitz, Layered representations for learning and inferring office activity from multiple sensory channels, Computer Vision and Image Understanding 96 (2) (2004) 163–180.
[14] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Conditional models for contextual human motion recognition, in: ICCV, 2005, pp. 1808–1815.
[15] S. Hongeng, R. Nevatia, F. Bremond, Video-based event recognition: activity representation and probabilistic recognition methods, Computer Vision and Image Understanding 96 (2) (2004) 129–162.
[16] P. Hong, M. Turk, T. Huang, Gesture modeling and recognition using finite state machines, in: Int. Conf. Automatic Face and Gesture Recognition, 2000, pp. 410–415.
[17] N. Ikizler, D. Forsyth, Searching video for complex activities with finite state models, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2007.
[18] R. Polana, R. Nelson, Detecting activities, in: IEEE Conf. on Computer Vision and Pattern Recognition, 1993, pp. 2–7.
[19] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.
[20] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: ICCV'03, 2003, pp. 726–733.
[21] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: ICCV, 2005, pp. 1395–1402.
[22] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: Int. Conf. on Computer Vision, 2007.
[23] I. Laptev, T. Lindeberg, Space-time interest points, in: ICCV, Washington, DC, USA, IEEE Computer Society, 2003, p. 432.
[24] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: ICPR, Washington, DC, USA, IEEE Computer Society, 2004, pp. 32–36.
[25] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: VS-PETS, October 2005.
[26] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, in: British Machine Vision Conference, 2006.
[27] S.F. Wong, T.K. Kim, R. Cipolla, Learning motion categories using both semantic and structural information, in: IEEE Conf. on Computer Vision and Pattern Recognition, June 2007.
[28] I. Laptev, P. Perez, Retrieving actions in movies, in: Int. Conf. on Computer Vision, 2007.
[29] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2008.
[30] I.N. Junejo, E. Dexter, I. Laptev, P. Perez, Cross-view action recognition from temporal self-similarities, in: European Conf. on Computer Vision, 2008.
[31] W. Freeman, M. Roth, Orientation histograms for hand gesture recognition, in: International Workshop on Automatic Face and Gesture Recognition, 1995.
[32] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2005, pp. I:886 – I:893.
[33] N. Ikizler, P. Duygulu, Human action recognition using distribution of oriented rectangular patches, in: Human Motion Workshop LNCS 4814, 2007, pp. 271–284.
[34] D. Ramanan, D. Forsyth, A. Zisserman, Strike a pose: Tracking people by finding stylized poses, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2005, pp. I:271 – I:278.
[35] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43 (1) (2001) 29–44.
[36] Y. Rubner, C. Tomasi, L.J. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Computer Vision 40 (2) (2000) 99–121.
[37] H. Ling, K. Okada, Diffusion distance for histogram comparison, in: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 246–253.
[38] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[39] J.C. Niebles, L. Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2007.
[40] L. Wang, D. Suter, Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model, in: IEEE Conf. on Computer Vision and Pattern Recognition, June 2007.
[41] Y. Ke, R. Sukthankar, M. Hebert, Spatio-temporal shape and flow correlation for action recognition, in: Visual Surveillance Workshop, 2007.