A Unified Framework for Locating and Recognizing Human Actions

Yuelei Xie[1,2] Hong Chang[1,2] Zhe Li[1,2] Luhong Liang[1,2] Xilin Chen[1,2] Debin Zhao[3]

[1]Institute of Computing Technology, Chinese Academy of Sciences (CAS), China

[2]Key Laboratory of Intelligent Information Processing, CAS, China

[3]School of Computer Science and Technology, Harbin Institute of Technology, China

{ylxie, hchang, zheli, lhliang, xlchen, dbzhao}@jdl.ac.cn

Abstract

In this paper, we present a pose based approach for locating and recognizing human actions in videos. In our method, human poses are detected and represented based on deformable part model. To our knowledge, this is the first work on exploring the effectiveness of deformable part models in combining human detection and pose estimation into action recognition. Comparing with previous methods, ours have three main advantages. First, our method does not rely on any assumption on video preprocessing quality, such as satisfactory foreground segmentation or reliable tracking; Second, we propose a novel compact representation for human pose which works together with human detection and can well represent the spatial and temporal structures inside an action; Third, with human detection taken into consideration in our framework, our method has the ability to locate and recognize multiple actions in the same scene. Experiments on benchmark datasets and recorded cluttered videos verified the efficacy of our method.

1. Introduction

Human action recognition is a challenging and widely studied problem in computer vision and pattern recognition community. Effective solutions to this problem can serve a lot of important application domains, such as human-computer interaction, security surveillance, the development of intelligent environments, etc. During the past decades, lots of approaches have been proposed to solve this problem. We can categorize the approaches into three major classes. One class of methods tries to model actions using global spatiotemporal templates, such as Motion-Energy-Image and Motion-History-Image templates proposed by Bobick and Davis [1] and space-time shape models by Blank et al. [2]. In [3], Efros et al. derive a robust motion descriptor from optical flow for identifying the instances of different actions. More recently, Yao and Zhu [4] present deformable action templates to detect interesting actions in videos.

The second class is based on space-time (S-T) interest point detectors [5, 6]. In [5], Niebles et al. recognize actions by applying latent topic model with a "bag of spatial-temporal words" representation for video sequence. Laptev et al. [6] present a new method for video classification by incorporating space-time pyramids and multi-channel non-linear SVMs with local S-T features. Although the sparsity and computational efficiency are appealing, the S-T points are hard to be reliably extracted from raw videos.

Both classes of approaches neglect human poses which convey the most important information of actions. Actually, an action can be seen as a temporal sequence of poses, and some key poses are usually fairly distinguishable from others of different actions. A reliable representation of pose can facilitate modeling complex and continuous actions. Recently, some researchers study action recognition based on informative pose descriptors. We refer to this class as pose based approaches. Examples to this class include [7, 8, 9]. In [7], Ikizler and Duygulu represent each human pose by spatial histogram of oriented rectangular patches extracted over the human silhouettes. Wang et al. [8] apply R transform to silhouettes to form a geometric invariant representation. Thurau and Hlavac [9] propose a HOG-based pose descriptor by exploiting a non-negative matrix factorization (NMF) basis representation. Our method also falls into this category.

Despite the promising performance of recent action recognition methods, some limitations of these solutions still exist. Firstly, many approaches assume that humans can be reliably tracked by some bounding boxes [10, 11]. Besides, approaches based on silhouette extraction assume that foreground can be nicely segmented from background. Both assumptions on the preprocessing step are critical for the following action recognition task. However, they are unrealistic and too hard, especially in some uncontrolled environment. Secondly, most approaches cannot be generalized to recognize multiple actions by different humans in the same scene.

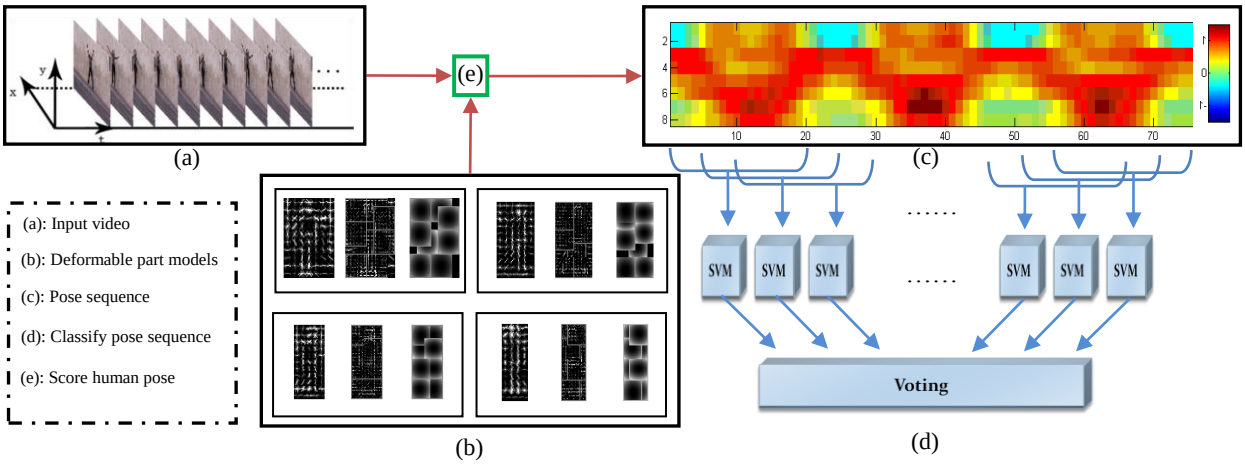

Figure 1: Framework of our action recognition algorithm. (a) input videos; (b) a set of learned deformable part templates; (c) visualization of pose sequence representations, where the horizontal axis corresponds to time step and the vertical axis corresponds to mixture components; (d) action classification by SVMs and majority voting.

In this paper, we try to make up the above limitations and propose a novel pose based approach to recognize actions by jointly considering three natural issues: human detection, pose estimation and action recognition. It is worth noting that there are only a few methods jointly considering the problems like this in the literature. In this way, not only can we weaken the assumptions like uncluttered background or reliable tracking in preprocessing, but also we can locate human actions in videos in both space and time dimensions. Meanwhile, it is straightforward for our method to simultaneously recognize multiple actions performed by different humans in the same scene. To illustrate the efficacy of our method, we conduct a series of experiments on both benchmark databases and recorded videos with multiple actions in cluttered conditions.

1.1. Overall Proposals

The overall framework of our pose based action recognition method is illustrated in Figure 1. Given some input videos (a), our method recognizes the actions inside each video according to two main steps. First, a set of deformable part templates (b) are applied to detect humans in the video frames following [12]. Instead of estimating the accurate pose configuration of body parts as in [14, 15], we just compute the matching scores of each detected human with respect to all templates. The scores are then concatenated to form our compact representation for a human pose. Thus, each action in the input videos is described using a sequence of pose representations (c), which reflects the spatial and temporal structures of different actions. Second, we classify the sub-sequences of these pose sequences by standard SVM, and assign an action class label to each whole sequence by majority voting, as shown in (d).

The main contributions of this work are three-fold:

1) We weaken the assumptions on the video preprocessing quality for action recognition
2) We propose a unified framework which combines human detection and pose estimation to solve the problem of action recognition. Particularly, the compact representation of human poses is proved effective and discriminating.
3) Our method has the ability to locate and recognize multiple human actions in the same scene.

The paper is organized as follows. Next section discusses the most related work to this paper. In Section 3, we describe the method for human detection and compact pose representation. Section 4 presents the action recognition algorithm. Experimental results are shown in Section 5. Section 6 concludes this paper.

2. Related Work

The pose based action recognition method proposed in this paper is inspired by related previous work in a few lines of research, although they are somewhat different from our method.

There has been only a few works in the literature that consider the problem of action recognition together with human detection and pose estimation. Examples include [16, 17, 18]. In [16], Yang et al. estimate body part locations using different standard linear SVM classifiers and classify

human action in an integrated way. Different from what we intend to solve here, their method focuses on action recognition from still images, under the assumption that there is only one person centered in the middle of every image. In [17], Ning et al. try to bridge the gap between high dimensional observations and random fields, by jointly optimizing the parameters of a latent pose estimator and random fields; however they assume that human has been detected while we try to integrate this step into action recognition. Our method is more similar in spirit with [18], where Ramanan and Forsyth annotate actions in videos by first tracking human then estimating 3D pose configuration and matching the pose to an annotated motion capture dataset. Different from their method, our method does not need to restore the accurate 2D configuration of body parts, thus, we can save computational complexity and avoid large amounts of manually labeling work.

Our method for action recognition is largely inspired by the success of deformation part model (DPM) in both human detection and pose estimation [12, 13, 14, 15]. The effectiveness of DPM implies great potentials to integrate the two problems into action recognition, which is also our ultimate goal. In [12, 13], Felzenszwalb et al. achieve state-of-the-art performance in object detection based on mixtures of deformable part models and histograms of oriented gradients (HOG) feature pyramid [19]. Ramanan et al. [14, 15] train deformable part models with pre-labeled images to restore accurate configuration of body parts. To our knowledge, our work is the first exploration on the effectiveness of DPM in combining human detection and pose estimation into action recognition.

As a feature representation method, HOG has shown its success in both human detection [19] and pose representation [9]. It also has the advantages of avoiding background subtraction in silhouettes based methods and distracting motion in dynamic feature based methods. In this paper, we adopt HOG pyramid to represent features.

3. Pose Representation

For pose based action recognition, it is crucial to derive a proper representation of human pose. It is straightforward to represent poses as in pose estimation literature [14, 15], but it usually requires large amounts of tedious manual work to label body parts in training images. In this paper, instead of restoring parts configuration, we match each human to a mixture of deformable part models and collect the matching scores for representing body pose.

Deformation part model, built upon pictorial structure [20], represents object by a collection of parts connected in a deformation manner. It explicitly encodes both appearance and spatial arrangements of different parts. DPM has been formulated as latent SVM and conditional random fields in human detection [12] and pose estimation [14] respectively. Based on [12], we integrate human detection and pose estimation into action recognition and avoid the large amounts of labeling work in pose estimation. In this section, we first briefly review the deformable part models proposed in [12], and then we extend their models to represent human poses.

3.1. Deformable Part Models

A deformable part model with n parts is defined as $M = (F_0, P_1, \ldots, P_n, b)$, where F_0 is a root filter, P_i is the model for the i-th part, and b is a real-valued bias term. Each part model P_i is defined as (F_i, d_i, v_i) where F_i is a filter for the i-th part, v_i is a two dimensional vector specifying a reference position for part i relative to the root position, and d_i is a two dimensional vector to penalize spatial displacement of part i with respect to its reference position. Let H denote the HOG feature pyramid of an image, and $p_i = (x_i, y_i, l_i)$ specify a position (x_i, y_i) in the l_i-th level of H for filter i. A configuration of filters can be denoted by $z = (p_0, p_1, \ldots, p_n)$. The matching score of configuration z to a deformable part model M depends on the responses of the filters at their respective locations, the deformation cost of each part with respect to the root, and the bias, i.e.,

$$score(z; M) = \sum_{i=0}^{n} F_i \cdot \varphi(H, p_i) - \sum_{i=1}^{n} d_i \cdot (dx_i^2, dy_i^2) + b \quad (1)$$

where $\varphi(H, p_i)$ is the sub-window of H covered by filter F_i with top-left corner at p_i, and (dx_i^2, dy_i^2) denotes a two dimensional vector containing the squared distances in x and y directions between the actual position of i-th part and its relative reference position to the root. As the level of each part is restricted to be twice the level of root in order to obtain high performance, the displacement is,

$$(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i) \quad (2)$$

At each root position in an image, a matching score is computed by maximizing over all possible placements of parts,

$$score(p_0; M) = \max_{p_1, \ldots, p_n} score(p_0, \ldots, p_n; M) \quad (3)$$

which can be obtained efficiently by generalized distance transforms [21]. High-scoring root locations indicate detected humans.

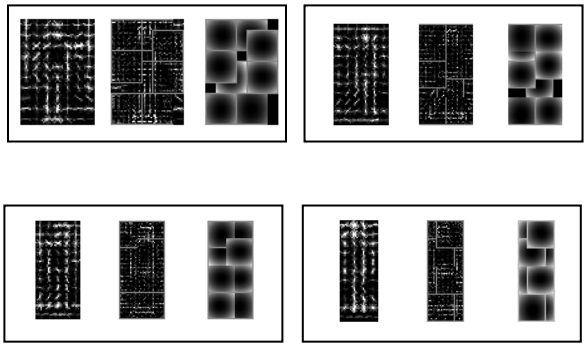

Figure 2: The 4-component mixture model used for detecting and scoring human poses. Each component contains a root filter, a set of part filters and deformable cost of each part, from left to right in a rectangle.

3.2. Training Deformable Part Models

The parameters of deformable part model can be trained in latent SVM framework where locations of part filter are treated as latent variables. Model parameters are trained by optimizing a discriminative function using stochastic gradient descent. It's recommended for interesting readers to refer to original work [12]. Here, due to space limitations we only focus on critical issues for representing body pose.

The first important issue is the initialization of part models. Unlike [15, 16], we only require human is labeled with a bounding box in training image, which helps save large amounts of labeling work. To initialize part models, an initial root filter is first trained using a standard SVM as in [19], then interpolated to twice the spatial resolution and a number of rectangles are greedily placed on the interpolated root filter to cover as much energy as possible. Part models are initialized by the sub-windows of the interpolated root filter covered by these rectangles. Intuitively, parts defined here do not explicitly correspond to physical parts of human body, like arms, legs, etc. but this method can help capture the most respective local appearance of human body.

Using only one deformable part model is not enough to capture significant variations in human appearance and pose articulation. As in [12] a mixture model is introduced to deal with this problem. Let $M = (M_1, M_2, ..., M_m)$ be a mixture model of *m* components. During detection, each root location in image is scored by each component in both left and right orientations, and the maximum score is used to define detections.

Mixture of deformable models can also be trained in the framework of latent SVM where the index of component giving the highest score is also treated as latent variable. In practice, we collected a set of 1841 images of different human poses as positive samples, some of the images are from Weizmann and KTH and some are from our daily life. We manually labeled humans in images with bounding boxes. Another set of background images is collected as negative samples. We train a 4-components mixture model on these images, which is shown in Figure 2.

3.3. Pose Representation Based on Multiple DPMs

Note that, once a human is detected, we have computed the matching scores of the detection with respect to all components. These scores can be interpreted as similarities between the body pose and different components. We propose to constitute a vector of the matching scores for pose representation. Suppose a person has been detected at position *p*, we concatenate the matching scores of all model components at *p* to form our representation. Let $score(p; M_c)$ denote the matching result of component M_c at *p*. As for mixture model, $score(p; M_c)$ is a vector $(score_0, score_1)$ with its elements corresponding to matching scores in left and right orientations respectively. Thus, our pose representation for a detection at position *p* can be expressed as,

$$pose(p) = (score(p; M_1), ..., score(p; M_m)) \quad (4)$$

Note that once a human is detected, the scores of different model components have been computed, and we immediately get the representation of human pose. Thus, different from previous methods, we derive the representation of human pose together with human detection.

4. Action Recognition

Representing actions: After we get the pose representations for a human at all time steps, an action can then be described by sequencing these pose representations in time order. Figure 3 shows example actions represented by our method with each corresponding to one action in Weizmann dataset. Columns in each pose sequence correspond to time steps, and rows correspond to components of mixture model. From these pose sequences, it's easy to observe clear temporal patterns which implies great potentials of our method for action recognition. In practice, in order to reduce the impact of imperfect detections, the pose sequence for an action is smoothed by averaging over a local window. A window of length 3 is chosen in our case.

SVM classification: As in [7, 9], the action recognition is performed by a windowing approach. The pose sequence is first segmented into fixed length sub-sequences, with neighboring sub-sequences overlapped by some ratio. We trained SVM classifiers with RBF kernel for each action respectively. The parameters of SVM classifier were

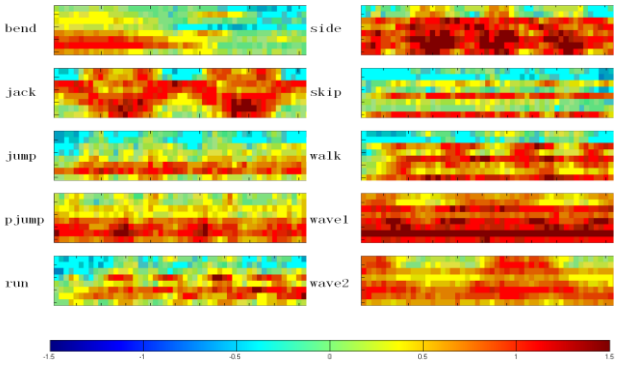

Figure 3: Examples of human actions represented by our pose representation method.

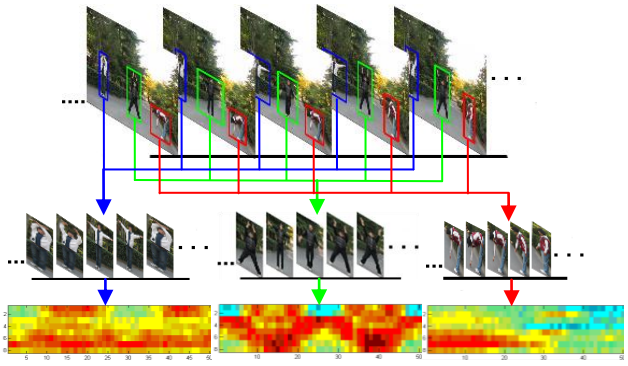

Figure 4: Detections corresponding to the same actor are grouped together.

selected by grid search and 10-fold cross-validation. These SVM classifiers are used to classify each sub-sequence, and the whole sequence is classified by a majority voting manner based on the predicted class labels of its sub-sequences.

Recognizing multiple actions: In practice, in order to recognize multiple actions in the same scene, we need to group these detections according to which person a detection corresponds to. The poses belonging to the same group are arranged in time order, forming the pose sequences of one person. In this paper, we use an online approach to group detections based on color histogram. A color histogram descriptor is maintained for each group and updated over time. A detection is grouped according to the smallest L1 distance between color histograms of the detection and any group. Let I_i (i=0,1,2,…) be the i-th frame in a video and *DetectHuman*(I_i) the procedure which returns the detected humans' bounding boxes and pose representations. Let *GL* denote the group list and *g* be an element in *GL*, the latest detection in *g* is denoted as *head(g)*. And the color histogram for a group *g* and a detection *d* are denoted as *chist(g)* and *chist(d)* respectively. Then our online grouping procedure works as follows:

1 Initialize *GL* with frame I_0
2 for each frame I_i, i = 1,2,3,…
3 *Det = DetectHuman(Ii)*
4 for each detection *d* in *Det*
5 Selecte the group *g* from *GL* that has the most similar color histogram with *d*:
mindist = L1Dist(chist(g), chist(d))
6 if *mindist > threshold*
7 add a new group *g'* initialized with *d* to *GL*
8 set *head(g') = d*
9 else if *head(g)* is from *Det*
10 if *mindist < L1Dist* (*chist(head(g)), chist(g)*)
11 set *head(g) = d*
12 else
13 set *head(g) = d*
14 for each group *g* in *GL*
15 if new detection was added to *g*
16 *chist(g)* = 0.90 * *chist(head(g))* + 0.10 * *chist(g)*
17 Return groups in *GL* that is longer than *L*

Where step 5 is for the situation when a new person is detected and step 9 is for when multiple detections for a same person are detected. The color histogram is updated by a weighting manner to account for all past detections in a group as in step 16. And in step 17 only groups that are longer than *L* are returned, this is because false detection could happen, and they usually form a shorter group due to step 6. *L* is set to be 25 in our experiment. Figure 4 shows an example of grouping detections for multiple actors in the same scene.

5. Experiments

We conduct three experiments to verify the efficacy of our approach. Two are carried out on the benchmark databases Weizmann [2] and KTH [23], and the other is on the database we recorded to validate the ability to recognize multiple actions.

Experiment 1: This experiment is evaluated on the Weizmann dataset which was originally recorded by Blank et al. [2]; it consists of 9 subjects with each performing a set of 10 different actions: bending down, jumping jack, jumping, jumping in place, skipping on one leg, galloping sideways, running, walking, waving one hand, and waving both hands. As most methods do with the dataset, we adopt a leave-one-out cross-validation scheme for testing: videos of 8 subjects are used for training, and the remaining one for testing, which is repeated for all 9 subjects and the results are averaged. In our windowing approach, each video is

segmented into sub-sequences of length 25 and neighboring

	5 parts	6 parts	7 parts	8 parts
3 components	78.89%	83.33%	85.56%	87.78%
4 components	84.44%	89.31%	93.33%	**95.56%**

Table 1: Precision of classifying action sequences under different numbers of parts and components.

sub-sequences have an overlapping of 3 frames. All training sequences are also flipped a same direction to make our method can classify the same actions with different facing orientations to the same class.

We first conduct a series of experiments on this dataset to explore the impact of the number of components in mixture model and the number of parts in each component on the overall performance. The results on classifying action sequences are summarized in Table 1. As we can see, the performance is increased with the number of components and parts, which is mainly because increasing the number of components and parts will improve the mixture model's ability to capture more local details. However, the more complex the model, the more computational time and danger of over fitting. For our task, it's reasonable and enough to use 4 components and 8 parts in the model, which gives the best result here.

The resulting confusion matrices are shown in Figure 5 and Figure 6. It can be seen that the "waving one hand" action is the most difficult one to be recognized and rather easy to be confused with the "waving two hands" action. This may be because instead of restoring body parts configuration, we derived our pose representation only based on the matching scores of each component and the contribution of part filters anchored at arm positions is relative small.

Table 2 shows the comparison results of different approaches evaluated on the Weizmann dataset. Note that most methods are performed in a "leave-one-actor-out" scheme with slight differences. As we can see, our method is comparable to state-of-the-art methods. However, our method do not need foreground segmentation as those based on silhouettes [2, 7] and we do not assume clean background in either training or testing phases as in [9]. What's more, we should emphasize that our methods integrate human detection and pose estimation into action recognition, which makes it possible to simultaneously recognize multiple actions in the same scene.

Experiment 2: The second experiment is performed on more challenging KTH dataset [23]. There are six actions in this dataset: boxing, handclapping, handwaving, jogging, running and walking. The actions are performed by 25

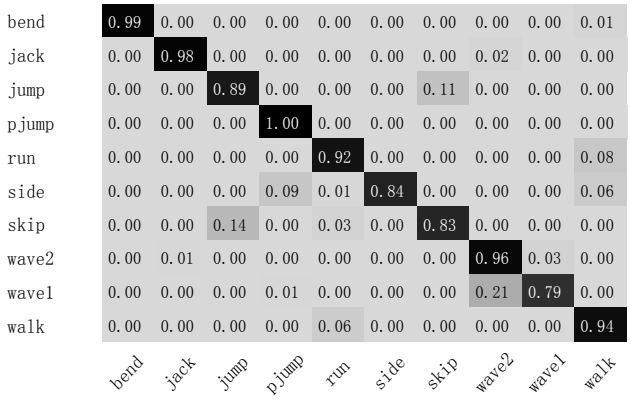

Figure 5: Confusion matrix for classifying subsequences on Weizmann dataset.

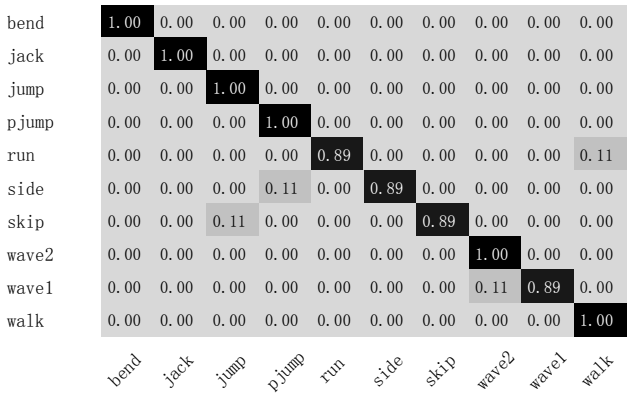

Figure 6: Confusion matrix for classifying action videos on Weizmann dataset.

Methods	Accuracy
Our Method	95.60%
Thurau [9]	94.40%
Ikizler [7]	100.00%
Blank [2]	99.60%
Niebles [5]	90.00%
Ali [10]	95.75%

Table 2: Comparison of Different approaches evaluated on Weizmann dataset

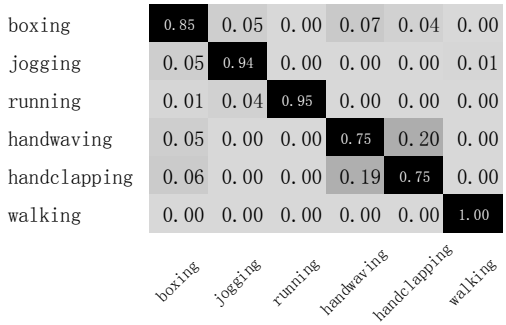

Figure 7: Confusion matrix for classifying actions on KTH dataset.

	all	d1	d2	d3	d4
Our method	87.3%	90.5%	75.4%	84.8%	93.7%
Yao [4]	87.8%	90.1%	84.5%	86.1%	91.3%
Ikizler [7]	89.4%				
Neibles [5]	81.5%				
Ali [10]	87.7%				

Table 3: Comparison of different approaches evaluated on KTH dataset

subjects under four different conditions (d1-d4) (see [23] for details). We use a similar experiment setting as in [4]. All the sequences are trimmed to 20 frames and flipped a same direction. The evaluation is performed with 5-fold cross-validations: the dataset is split into 5 folds with 5 subjects in each, 4 folds for training and one for testing. The results are averaged over 5 permutations.

Figure 7 shows the results of our method on KTH dataset. Although jogging and running are the hardest actions to be distinguished in the dataset, our method distinguishes the two actions very well. Actually, the poses involved in jogging and running are very similar, while the speeds of the pose evolutions in the two actions are somewhat different from each other, which can be used as the critical information to distinguish actions. Since we represent a human action by a sequence of poses, the speed information has been implicitly encoded in our representation. Thus our method can successfully recognize the two actions. This is also one advantage of our pose based approach. Moreover, we test our methods under four different conditions respectively. Table 3 summarizes the comparison results, where our method achieves comparable performance to state-of-the-art methods.

Experiment 3: In this experiment, we validate our method in recognizing multiple actions in the same cluttered scene. The experiment is conducted on a collection of videos recorded by us. Each video contains more than one actors who are performing different actions in a relative complex setting. The videos involve five action classes selected from Weizmann dataset: bending down, jumping jack, jumping

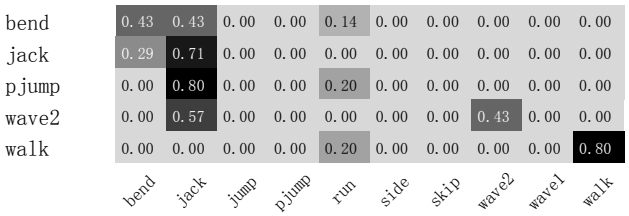

Figure 9: Results for classifying recorded action videos against models trained on the Weizmann dataset. We achieved the averaged precision 47.3%.

in place, walking, and waving both hands. Some example frames in the recorded video are shown in Figure 8. To recognize the actions inside the video, we first detect humans in videos frames, extract pose representation for each detection, and then these detections are grouped to form pose sequences for each person's action using the method described in section 4. The ground truth action class for pose sequence is labeled manually. We predict the action class of pose sequences against SVM classifier trained on Weizmann dataset and the result is averaged. Figure 9 shows the confusion matrix for classifying pose sequences and some detection results are shown in Figure 10.

There are mainly two reasons that we did not achieve perfect results as in previous two experiments. First, models trained on Weizmann dataset are not robust enough to cluttered background. Second, for multiple actions recognition, it could happen at some time steps two actions are overlapped with each other. More reliable methods for handling these two problems will be our future work.

6. Conclusion

In this paper, we proposed a method for compactly representing human pose to recognize human actions in videos, which is based on deformable part models and works together with human detection. An action is represented by a sequence of pose representations on which action classification is performed. Comparing with previous methods, ours have three main advantages: a) we do not rely on foreground segmentation as those based on silhouettes; b) instead of representing body poses by body parts configuration as in traditional methods of pose estimation, we derive a more compact representation for body pose, in addition, only bounding boxes of the persons in training images are needed, while in traditional pose estimation one has to label the accurate configuration of body parts; c) As we take human detection into consideration, our methods have the ability to locate and recognize multiple actions in the same scene. However, there still exist limitations in our current work. One is that we need two training procedures, one for mixture model and the other for SVM classifier. The

Figure 8: Example frames of our recorded videos

Figure 10: Example frames for detecting multiple humans in the same scene.

other is that temporal action segmentation is assumed as the datasets we evaluated on, actually the voting scheme we adopt here can also do temporal segmentation, however more robust scheme based on our pose descriptor will be one of our main future topics.

Acknowledgment

This paper is partially supported by Natural Science Foundation of China under contracts No. 61025010, No. 60803084, and No. 60832004; National Basic Research Program of China (973 Program) under contract 2009CB320902.

References

[1] A. Bobick and J. Davis. The recognition of human movement using temporal templates. In *T-PAMI*, 23(3): 257-267, 2001.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *T-PAMI*, 2:1395-1402, 2007.

[3] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[4] Benjamin. Yao and Song-Chun Zhu. Learning deformable action templates from cluttered video. In *ICCV*, 2009.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *IJCV*, 79(3): 299-318, 2008.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[7] N. Ikizler and P. Duygulu. Histogram of oriented rectangles: A new pose descriptor for human action recognition. In *IVC*, 27(10): 1515-1526, 2009.

[8] Ying Wang, Kaiqi Huang, Tieniu Tan. Human activity recognition based on R transform. In *CVPR*, 2007.

[9] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.

[10] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. In *T-PAMI*, 32(2): 288-303, 2010.

[11] Yang Wang and Greg Mori. Learning a discriminative hidden part model for human action recognition. In *NIPS*, 2008.

[12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. In *T-PAMI*, 32(9): 1627-1645, 2010.

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[14] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[15] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *CVPR*, 2006.

[16] Weilong Yang, Yang Wang, and Greg Mori. Recognizing human actions from still images with latent poses, In *CVPR*, 2010.

[17] Huazhong Ning, Wei Xu, Yihong Gong, and Thomas Huang. Latent pose estimator for continuous action recognition, In *ECCV*, 2008.

[18] D. Ramanan and D.A. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.

[19] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[20] M.A. Fischler and R.A Elschlager. The representation and matching of pictorial structures. In T-C, 22(1): 67-92, 1973.

[21] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. In *IJCV*, 61(1): 55-79, 2005.

[22] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VSPETS*, 2005.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.