Pose Sentences:
A new representation for action recognition using sequence of pose words

Kardelen Hatun, Pınar Duygulu
Bilkent University, Department of Computer Engineering, Ankara, Turkey
$\{$*kardelen, duygulu*$\}$*@cs.bilkent.edu.tr*

Abstract

We propose a method for recognizing human actions in videos. Inspired from the recent bag-of-words approaches, we represent actions as documents consisting of words, where a word refers to the pose in a frame. Histogram of oriented gradients (HOG) features are used to describe poses, which are then vector quantized to obtain pose-words. As an alternative to bag-of-words approaches, that only represent actions as a collection of words by discarding the temporal characteristics of actions, we represent videos as ordered sequence of pose-words, that is as pose sentences. Then, string matching techniques are exploited to find the similarity of two action sequences. In the experiments, performed on data set of Blank et al., 92% performance is obtained.

1 Introduction

Recognition of human actions is a well-studied yet still challenging problem (see [6, 7, 10] for recent surveys). Representation of actions is an important factor for recognition. In some group of studies the entire action sequence is combined into a single spatio-temporal representation [1, 2], while in another group, the actions are represented in the form of basic action units or action primitives [5, 9].

In recent studies, the bag-of-words approaches, inspired from text and used for object and scene recognition in computer vision, are also applied to recognize actions as an alternative form of descriptive action units. In these approaches, actions are represented as a collection of visual words which are the codebooks of spatio-temporal features. Examples include the space-time interest points used in an SVM based method by Schuldt et al. [14], histogram of cuboids by Dollar et al. [4], the pLSA approach applied on cuboids by Niebles et al. [12], and histogram of rectangles approach in [8].

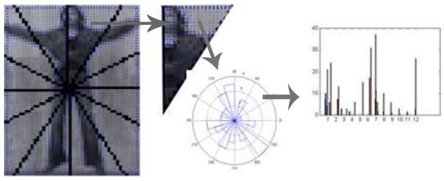


Figure 1. HOG feature extraction.$(n, m = 12)$

In this study, we propose a new representation for recognition of human actions. Following the idea of bag-of-words approaches, and considering the pose as an important factor for understanding of actions, we describe the poses in each frame of an action sequence as visual words, which we refer as **pose-words**.

To represent the pose in each action frame, we use the histogram of gradients (HOG) approach [3] which was originally proposed to localize humans in images. Pose-words are then constructed by vector quantization of HOG features extracted from each frame.

Our main contribution lies in the use of visual words. Unlike the bag-of-words approaches that represent the actions only as a collection of visual words, by discarding the temporal information which is an important characteristics of actions, we represent the actions as ordered sequence of pose-words, that is in the form of **pose-sentences**. We then propose a method to match the actions using string matching techniques.

2 Related Work

In [16], similar to our approach, Wang et al. code a frame in an action sequence with a single word unlike the other approaches representing it as a collection of spatio-temporal codebooks. However, they use a semi latent Drichlet allocation approach to represent actions as a bag of coded frames, discarding the temporal information. Also, they represent the frames using the motion descriptor obtained from optical flow vectors, while we use HOG to capture the shape of the pose.

Preprint submitted to 19th International Conference on Pattern Recognition.
Received April 15, 2008.

In [15], Thurau used HOG to define action primitives. Then, action recognition is considered as a sequence comparison problem, and n-grams are exploited for this purpose.

As an another study which aims to capture the temporal order of features, Nowozin et al. [13] represent a video as a sequence of sets of discretized spatio-temporal words, and use discriminative subsequence mining algorithm for classifying actions.

3 Our approach

In our approach, each action sequence A_i in the data set is represented as an ordered sequence of pose-words. To obtain pose-words, first, poses in each frame $f_{ij} \in A_i, j = 1 \dots |A_i|$ are described using the histogram of oriented gradients (HOG) features obtained from a radial partitioning of the frame. Then, all the frames in the data set are grouped according to their similarities to obtain pose-clusters. The centroids of each cluster are then defined as pose-words, $P = \{p_1 \dots p_K\}$. Then, an action sequence is coded in the following manner. If a frame in an action sequence belongs to cluster with a centroid p_k, then the frame is coded with the pose-word p_k. As an ordered sequence of pose-words, a pose-sentence representing the action with length N is then described as $A_i = a_1 a_2 \dots a_N$, where each a_n corresponds to a pose-word $p_k \in P$. Finally, string matching techniques are used to find the similarity of two pose-sentences. In the following, the steps of the method will be described in detail.

3.1 Feature extraction

To describe each frame, we use histogram of oriented gradients (HOG) feature, which was proposed in [3] for human detection in videos. Figure 1 summarizes the feature extraction process.

In the first step, the gradients in a frame f_{ij} are obtained by applying the 1-D $[-1\ 0\ 1]$ filter (which is shown to be best in [3]) in both x and y directions on the graylevel image of the frame to obtain G_x and G_y for each pixel.

Then, each frame is divided into n cells using a radial (circular) grid structure. In each cell, for m directions over the interval $[0, 2\pi]$, the gradient magnitudes of the pixels in that direction are summed, to obtain the HOG feature. Then, n histograms are attached to each other to obtain a $n \times m$ length feature vector for each frame, describing the shape of the pose.

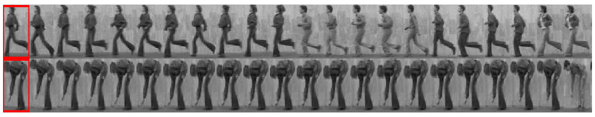


Figure 2. Samples from some clusters are shown, with the cluster centroids in red.

3.2 Generation of pose-words

To generate the pose words, the following steps are applied. First we form a similarity matrix S, where S_{ij} is the similarity of frame i and frame j. Then, we apply k-medoids algorithm on S to obtain K clusters. Finally, to build the codebook $P = \{p_1 \dots p_K\}$, the centroid of each cluster is taken.

As shown in Figure 2, the clusters are mostly coherent, corresponding to the same action, performed with the same or different actors, with minor problems.

3.3 Action recognition using Pose-words

The next step is to match the action sequences represented as ordered sequence of pose-sentences. Here, we first explain the **bag-of-poses** method, a simple method to represent the actions as a collection of pose-words as in the bag-of-words approaches; then we present our method based on **pose-sentences** where we capture the temporal order of the pose-words. In both cases classification of actions are performed using the nearest neighbor classifier with leave one out cross validation method.

3.3.1 Bag-of-poses method

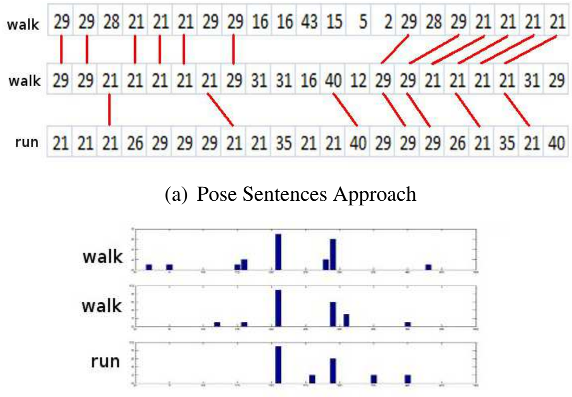
To simulate the bag-of-words approaches in the simplest way, we represent the action sequences as histograms of pose-words. Let A_i be an action sequence and K be the number of pose words. In the bag-of-poses method, we represent A_i by a $1 \times K$ bins histogram $h_1 \dots h_K$, where each bin h_k corresponds to the number of frames represented with the pose word p_k.

The similarity between two action sequences A_i and A_j is then defined using the Chi-square distance as

$$\chi^2(A_i, A_j) = \frac{1}{2} \sum_n \frac{(A_i(n) - A_j(n))^2}{A_i(n) + A_j(n)} \qquad (1)$$

3.3.2 String matching on pose-sentences

In order to capture the temporal characteristics of actions, we represent the actions in the form of ordered sequences rather than simply using bag-of-poses. That is we represent an action A_i as a pose sentence



(a) Pose Sentences Approach

(b) Bag-of-Poses Approach

Figure 3. These sequences can be discriminated with pose-sentences approach.

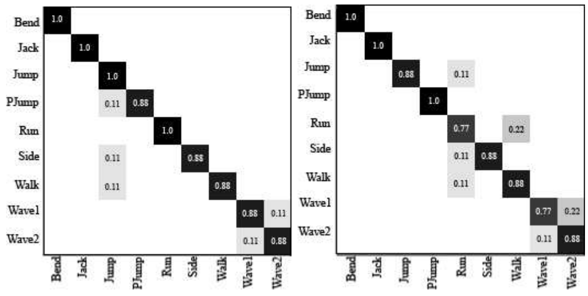
$a_1a_2 \dots a_N$, where $N = |A_i|$ and each a_n is a pose-word $p_k \in P$.

To find the similarity of two actions A_i and A_j represented in the form of pose-sentences, we then use a very simple string matching algorithm, *edit distance*[11]. With the edit distance algorithm, distance between two strings is defined as the minimum number of steps to be taken to convert A_i to A_j.

To understand the advantages of string matching approach over bag-of-words approach, let's consider the example given in Figure 3. These partial sequences are taken from two examples of walk actions, and one example of run action. The numbers represent the pose-words describing the frames in the sequence. When we consider the distribution of pose-words, we observe that pose-words 29 and 21 are representatives both for walk and run actions. While the bag-of-poses approach, which counts the number of occurrences of each pose-word in the sequence, captures the similarities between the two walk actions in general, the second walk action is more similar to the run action than to the first walk action, and therefore it is likely that it will be misclassified. On the other hand, our pose-sentence based approach encodes the ordering information, and therefore makes two walk sequences to be more similar compared to the run action.

4 Experiments

The experiments are carried out on the data set of Blank et al. [1]. This dataset consists of nine actions, `jump in place`, `wave one hand`, `walk`, `jack`, `bend`, `wave both hands`, `run`, `side` and `jump forward`, performed by nine people . There are a total of 81 videos and 5098 frames.



(a) Confusion matrix for pose sentences approach

(b) Confusion matrix for bag-of-poses approach

Figure 4. Confusion matrices for two classification methods

In order to construct the pose-words, first we performed a slightly controlled experiment. We handpicked 47 poses, which best represent and discriminate the actions, as the initial centroids. Then, we run the k-medoids method on these initial centroids, to obtain the final distribution of clusters.

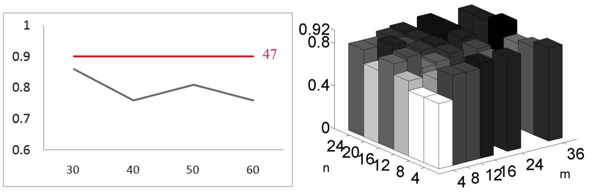
The actions are then represented in two forms: (a) as a bag-of-poses representation, (b) as a pose-sentence. For both representations, in order to perform the classification of actions, we use the leave one out cross validation scheme. We choose one example from one action as a test, and use the rest of the 80 examples as training. Then, we perform the nearest neighbor classification and label the test action with the label of the most similar action in the training set. To find the similarities, we use the Chi-square distance for bag-of-poses representation, and the edit distance for the pose-sentence representation.

Figure 4 shows the confusion matrices for the bag-of-poses and pose-sentence based approaches. As shown in the figure, pose-sentences based approach perfectly classifies 4 of the actions, while mis-classifies only one example in the remaining 5 actions. It confuses `wave one hand` with `wave two hands`, and `jump forward` with `jump in place` which are very similar actions. On the other hand, bag-of-poses approach produces more mis-classified results. Due to the missing temporal information, it cannot capture the differences between run and walk, unlike the pose-sentences approach.

We compare the overall success rate of our approach with the related studies experimented on the same data set. We see that, the pose-sentences approach is superior to the pLSA based approach which uses the spatio-temporal words [12], and to the n-gram based approach which also uses a HOG description for representing the words [15]. In this study, we use a very simple classification method to concentrate on the representation. The

Matching Method	Success Rate
Ikizler[8]	100%
Blank[1]	99%
Our Approach	**92%**
Bag-Of-Poses	88%
Thurau[15]	87%
Niebles[12]	73%

Table 1. Comparison with related studies



(a) Success rates for varying K values (b) Success rates for varying m and n values

Figure 5. Tests performed for pose sentences approach.

results promise that with a more complicated classification method, the results can be in a similar level with the best results in the literature.

As we mentioned, these results are obtained by fixing K to 47 and choosing hand-picked centroids for initialization of the k-medoids algorithm. In order to understand the choice of K in a randomly initialized k-medoids clustering algorithm, we choose K = 30, 40, 50, 60 values, and record the performance as shown in Figure 5(a) for fixed values m=24 and n=24. The results show that, although the choice of K affects the performance, the results are still in a similar level, and even with random initialization K around 50 is an acceptable choice.

In the extraction of HOG features, the choice for number of cells n, and number of orientation bins m is important. In order to test the effect of these parameters, we fix the number of centroids K=47, and run the algorithm for different n and m values as shown in Figure 5(b). The set of values tried for each parameter in each test are: $n = 4, 8, 12, 16, 20, 24$ and $m = 4, 8, 12, 16, 24, 36$ respectively. The results suggest higher values for m and n, and show that the orientation bin size is more important.

5 Conclusion

In this study, we propose a new method for representing actions in the form of ordered sequence of pose-words as an alternative to bag-of-words approaches which discard the temporal ordering. 92% performance on a benchmark data set, with a simple classification method, justifies the importance of the proposed representation. The proposed method combines the pose information with the temporal information. Simple HOG features are used to encode poses, and simple string matching techniques are used for encoding the temporal similarities. In the future, we plan to improve the results with adapting more complicated methods.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(3), March 2001.

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

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[5] P. Fihl, M. Holte, and T. Moeslund. Motion primitives for action recognition. In *Workshop on Gesture in Human-Computer Interaction and Simulation*, 2007.

[6] D. Forsyth, O.Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion i: Tracking and animation. *Foundations and Trends in Computer Graphics and Vision*, 1(2/3), 2006.

[7] W. Hu, T. Tan, L. Wang, and 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.

[8] N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectanguar patches. In *Human Motion Workshop, (with ICCV)*, 2007.

[9] O. Jenkins and M. Mataric. Deriving action and behavior primitives from human motion data. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2002.

[10] V. Kruger, D. Kragic, A. Ude, and C. Geib. The meaning of action: A review on action recognition and mapping. *Advanced Robotics*, 21(13):1473–1501, 2007.

[11] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966.

[12] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[13] S. Nowozin, G. Bakir, and K. Tsuda. Discriminative subsequence mining for action classification.

[14] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach.

[15] C. Thurau. Behavior histograms for action recognition and human detection. In *Human Motion Workshop, (with ICCV)*, 2007.

[16] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Human Motion Workshop, (with ICCV)*, 2007.