Shape Prototype Signatures for Action Recognition

Michael Donoser, Hayko Riemenschneider and Horst Bischof[*]
Institute for Computer Graphics and Vision, Graz University of Technology
$\{$*donoser,hayko,bischof*$\}$*@icg.tugraz.at*

Abstract

Recognizing human actions in video sequences is frequently based on analyzing the shape of the human silhouette as the main feature. In this paper we introduce a method for recognizing different actions by comparing signatures of similarities to pre-defined shape prototypes. In training, we build a vocabulary of shape prototypes by clustering a training set of human silhouettes and calculate prototype similarity signatures for all training videos. During testing a prototype signature is calculated for the test video and is aligned to each training signature by dynamic time warping. A simple voting scheme over the similarities to the training videos provides action classification results and temporal alignments to the training videos. Experimental evaluation on a reference data set demonstrates that state-of-the-art results are achieved.

1 Introduction

Action recognition is currently one of the most investigated fields of research in computer vision. It has many applications as e. g. in visual surveillance, content-based video search, human computer interaction or sports analysis. In general actions are defined by diverse features like appearance and shape as well as dynamic cues like spatio-temporal trajectories and optical flow fields. Combining these features in probabilistic settings has shown to be an effective way to improve action recognition performance [7].

Many of the recent state-of-the-art approaches intermingle several features for action recognition. Mostly, it is hard to judge which features contribute most to the final result. A recent trend suggests that a combination of motion features like optical flow and appearance features like the powerful HoG descriptor achieves best performance [13]. The question we seek to answer in this paper is how much can the shape of the human silhouette alone contribute to an action recognition system. This follows recent conclusions [11, 16] that action recognition can be performed even from single frames, by simply looking at the pose of the human which is mainly defined by the silhouette shape.

In this paper we address the problem of action recognition from videos by solely analyzing shape cues of the human silhouette. We assume a rather constrained, static video acquisition setting as e. g. it is common in many visual surveillance scenarios. This allows providing approximate figure/ground segmentation results for the humans in every frame. The key idea of our approach is that we describe an entire sequence by a signature of similarities to shape prototypes. Such an approach using similarities to prototypes was also proposed by Weinland et al. [16], but in contrast we use the similarities to all shape prototypes as discriminative descriptor for a video sequence, which is easily aligned to other sequences by dynamic time warping. We show that these shape signatures are a powerful descriptor, where a very simple voting scheme is sufficient to obtain state-of-the-art results on a reference data set.

2 Action recognition

Our method mainly consists of two subsequent steps. First, we identify shape prototypes from a set of provided human silhouettes containing the relevant poses of all actions to be recognized. These prototypes are found by a novel shape clustering method which is outlined in detail in Section 2.1. Second, for all videos in the training set, we compare the extracted human silhouettes to each of the C obtained prototypes by a shape matching method which yields an $C \times t$ similarity descriptor for a video sequence of t frames. This step is outlined in Section 2.2. Finally, to classify a test video sequence, we again build the shape prototype signature and align it to all videos of the training set by dynamic

[*]This work was supported by the Austrian Research Promotion Agency (FFG) project FIT-IT CityFit (815971/14472-GLE/ROD) and the Austrian Science Fund (FWF) under the doctoral program Confluence of Vision and Graphics W1209.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.443

time warping as it is described in Section 2.3. A simple voting scheme using the dynamic time warping distance is used to classify the sequence which yields state-of-the-art results on a reference data set as it is demonstrated in Section 3.

2.1 Shape prototype selection

As a first step, we automatically identify shape prototypes from a set of provided human silhouettes including all relevant poses from the actions to be recognized. For prototype identification we apply a shape matching based clustering approach. We first we build a pairwise shape similarity matrix by comparing all available shapes to each other and then apply a pairwise clustering method on the calculated matrix.

Shape matching is a well investigated problem in computer vision and several powerful methods exist [2]. In general, the applied method should be invariant to similarity transformations and robust against noise and outliers. Furthermore, since in our scenario an all vs. all comparison has to be performed, shape matching has to be as efficient as possible. We selected the COPAP method [12], since the code is publicly available and their method has shown to provide excellent results for closed contours (since point ordering is considered) as they are available in our setup. COPAP provides a similarity score between two binary input silhouettes. The additionally provided correspondences are not further considered.

We assume that we have given P human silhouettes and compare each of the shapes to all others. Since the similarity scores a_{ij} provided by COPAP are not symmetric, this requires N^2 comparisons and finally leads to an $N \times N$ affinity matrix $W = \{w_{11}, \ldots, w_{NN}\}$ which is obtained from the calculated distance matrix $A = \{a_{11}, \ldots, a_{NN}\}$ by

$$w_{ij} = \exp\left(-\frac{a_{ij}^2}{\sigma_{ij}^2}\right) . \qquad (1)$$

As outlined in several papers the choice of the normalization parameter σ_{ij} is important to achieve good clustering results. We normalize according to a method proposed in [17] which defines

$$\sigma_{ij} = \sigma_i \sigma_j \quad with \quad \sigma_i = A(i, i_K) \; , \qquad (2)$$

where i_K is the K'th nearest neighbor of shape i and K is a parameter fixed to 8 in all experiments.

To obtain a cluster result and prototypes per cluster it is possible to apply any pairwise clustering method to the obtained affinity matrix W. The shape similarity measure is not necessarily metric which has to be taken into account by the clustering method. Unfortunately, most classical methods for pairwise clustering as e. g. K-Means or spectral clustering only consider metric similarities. Obviously a pairwise clustering method considering non-metric spaces would be preferable.

Therefore, we selected a recently proposed method denoted as Affinity Propagation clustering [5] which showed impressive performance on several data sets. Affinity propagation is based on iteratively exchanging messages between the data points until a good solution emerges. While most clustering methods only keep track of some candidate exemplars during search, affinity propagation considers all data points as candidates. Furthermore, affinity propagation is also able to handle missing data points and non-metric similarity measures and it returns one of the data-points as prototype per cluster. Therefore, it is perfectly suited for shape clustering.

Thus, applying affinity propagation clustering onto our $N \times N$ affinity matrix W divides our training set into C different clusters and each cluster is represented by a single prototype P_c. The so-called preference parameter of affinity propagation furthermore allows to influence the number of prototypes. Please note, that in the experiments outlined in Section 3 we always use the default preference parameter, which is the median of the affinity matrix W.

2.2 Shape prototype signatures

After obtaining the prototypes P_c as described in the previous section we are now able to describe a video in terms of a signature of similarities to the shape prototypes. We again use COPAP as shape matching method and compare the obtained silhouette of every frame in the sequence to all of the prototypes P_c which yields an $C \times t$ shape prototype similarity signature for a sequence of t frames. This process is illustrated in Figure 1, where we show the temporal changes of the similarity scores to two selected prototypes for a bending action sequence. As can be seen, the scores smoothly approach and retreat high similarity values to the individual prototypes.

As it is illustrated in Figure 2 these signatures are highly discriminative. In the next section we demonstrate that even a very simple voting scheme enables to achieve state-of-the-art results on a reference data set.

2.3 Classification by signature alignment

To classify a given video sequence we calculate its prototype signature as described in the previous section and compare it to all signatures of our training data

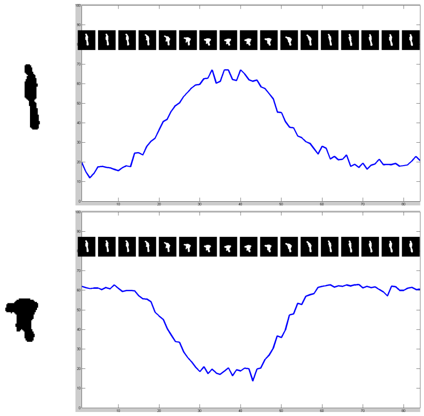

Figure 1. Illustration of prototype signature calculation for a bending action. In every frame shape differences to prototypes (here two are shown) are calculated. Low scores express high similarity.

set. To be able to compare two signatures we still have to cope with the fact that the sequences contain different numbers of frames. In [16] this problem was addressed by only considering an unordered set of prototype matches as video descriptor. Of course, in such an approach the entire temporal information between consecutive human poses is lost. In contrast, we use the entire shape signatures and align them to each other.

Alignment of time signals can be done by many different methods, where dynamic time warping (DTW) [10] is one of the most prominent ones. The idea behind dynamic time warping is to align two input signals (as e. g. our shape signatures) by warping their time axes and by finally measuring the similarity between the warped signals. In a first step, DTW calculates signal to signal similarities between all data points of the two signals. In our case we use Euclidean distances as measure, where each signal point is defined by the C-dimensional vector containing the similarities to each of the C prototypes.

DTW provides a score between the two input signatures. For classification, we calculate the score to all training videos and simply assign the label of the video with the highest score. Please note, that this simple voting approach yields state-of-the-art results on a reference data set as shown in the next section, but of course one can use more sophisticated approaches for this step, as e. g. learning the shape signatures by classifiers like SVM or the recently popular random forest classifier.

Table 1. State-of-the-art average recognition rates for Weizmann data set [6].

Lin et al. (shape only) [8]	81.1%
Ali et al. [1]	92.6%
Bregonzio et al.[3]	96.7%
Riemenschneider et al. [9]	96.7%
Weinland et al.(50 exemplars) [16]	97.7%
Wang and Suter [14]	97.8%
Lin et al. (shape+motion) [8]	100%
Wang and Mori [15]	100%
Fathi and Mori [4]	100%
Our method (21 prototypes)	**100%**

3 Experiments

To evaluate the performance of our proposed shape guided action recognition method, we applied it on the well-known Weizmann data set [1] [6]. This data set consists of ten different types of actions: bending, jumping jack, jumping, jump in place, running, side jumping, skipping, walking, one-hand and two-hand waving performed by nine different humans. Testing is performed in a leave-one-out-fashion on a per person basis, i. e. training is done on eight subjects and testing on the unused subject and all its videos. Analogue to recent papers, the average correct classification rate is calculated as final result.

Using the simple voting scheme presented in the previous section yields a perfect recognition result of 100% on the Weizmann data set. In comparison, Table 1 summarizes recent state-of-the-art results. As can be seen several authors already achieved a 100% recognition score, but using much more sophisticated approaches considering appearance, shape and motion cues. For example in [8] also a perfect recognition result was achieved analyzing motion and shape information, but solely based on shape data the reported score was only 81.1%. Our method only analyzes shape cues, neglecting any appearance and motion information in the videos, but nevertheless achieves a 100% recognition rate. Furthermore, we achieve this result by using only 21 shape prototypes, whereas the comparable method of [16] requires more than 100 prototypes, their score using 21 exemplars is only approximately 90%. To sum up, it seems that analyzing the shape of human silhouettes alone is sufficient to achieve reasonable action recognition results in such constrained scenarios.

[1] www.wisdom.weizmann.ac.il/˜vision/SpaceTimeActions.html

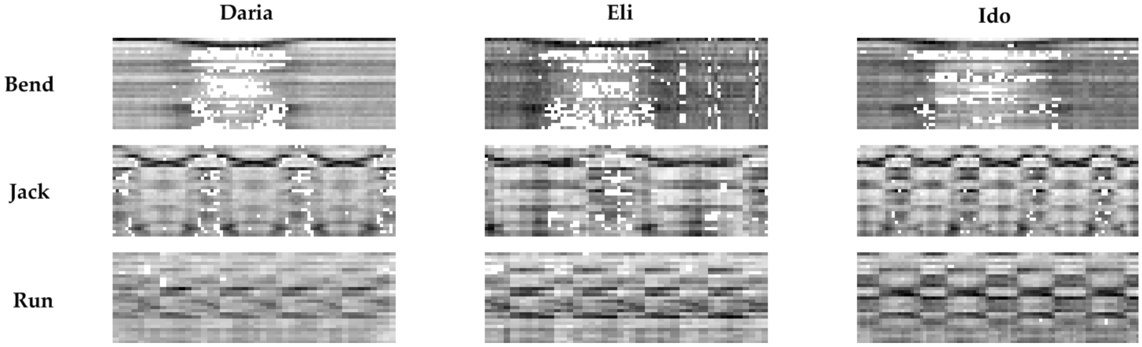

Figure 2. Prototype similarity signatures (21 prototype rows and 100 frame columns per descriptor) for three different actions performed by three humans. As can be seen the signatures to the 21 prototypes are highly discriminative but bear reasonable similarities between the same action performed by different humans.

4 Conclusions

This paper introduced signatures of similarity scores to pre-defined human silhouette prototypes as powerful descriptor for recognizing actions. In contrast to recent methods, we use similarity information to all prototypes by temporally aligning the descriptors of different video sequences. Using the signatures, we achieve state-of-the-art results on a reference data set by a very simple voting scheme. Results demonstrate that analysis of the shape of the human silhouette alone can significantly contribute to action recognition systems in constrained scenarios like a static video acquisition setting. Future work will focus on analyzing the performance on other data sets including more difficult scenarios using learning techniques to classify different shape prototype signatures.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Intern. Conf. on Computer Vision*, pages 1–8, 2007.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Trans. on Pattern Analysis and Machine Intel.*, 24(4):509–522, 2002.

[3] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *Conf. on Computer Vision and Pattern Recognition*, pages 1948–1955, 2009.

[4] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *Conf. on Computer Vision and Pattern Recognition*, 2008.

[5] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[6] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Trans. on Pattern Analysis and Machine Intel.*, 29:2247–2253, 2007.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conf. on Computer Vision and Pattern Recognition*, 2008.

[8] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *Intern. Conf. on Computer Vision*, 2009.

[9] H. Riemenschneider, M. Donoser, and H. Bischof. Bag of optical flow volumes for image sequence recognition. In *British Machine Vision Conf.*, 2009.

[10] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[11] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *Conf. on Computer Vision and Pattern Recognition*. IEEE Press, June 2008.

[12] C. Scott and R. Nowak. Robust contour matching via the order-preserving assignment problem. *IEEE Transactions on Image Processing*, 15:1831–1838, 2006.

[13] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conf.*, 2009.

[14] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *Conf. on Computer Vision and Pattern Recognition*, 2007.

[15] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *Conf. on Computer Vision and Pattern Recognition*, 2009.

[16] D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In *Conf. on Computer Vision and Pattern Recognition*, 2008.

[17] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2004.