

Encoding Actions via Quantized Vocabulary of Averaged Silhouettes

Liang Wang

*Department of Computer Science
University of Bath, United Kingdom, BA2 7AY*

Christopher Leckie

*Department of Computer Science and Software Engineering
University of Melbourne, Australia, Vic 3010*

Abstract

Human action recognition from video clips has received increasing attention in recent years. This paper proposes a simple yet effective method for the problem of action recognition. The method aims to encode human actions using the quantized vocabulary of averaged silhouettes that are derived from space-time windowed shapes and implicitly capture local temporal motion as well as global body shape. Experimental results on the publicly available Weizmann dataset have demonstrated that, despite its simplicity, our method is effective for recognizing actions, and is comparable to other state-of-the-art methods.

1. Introduction

Visual analysis of human movement [4] aims to detect, track and recognize people, and more generally, to understand human activities from video sequences. In recent years, human action recognition has received growing interest, which is driven by a wide range of promising applications such as smart surveillance, human-machine interaction and motion capture. However, recognizing actions remains a challenge due to variations in both the environment (which determines the quality of visual clues extracted from videos) and the motion itself (which exhibits spatial and temporal variations among subjects with different physical characteristics, motion styles and speeds).

Motion characterization is central to the interpretation of human actions. Various visual cues have been used, *e.g.*, trajectories [1], interest points [11, 12], optical flow [3, 8], silhouettes [2, 15, 9, 7], to name a few. Two major strategies have been studied for motion modeling and recognition, namely template-matching meth-

ods and state-space methods [4]. The former generally converts time-varying features into a static pattern (*i.e.*, template) for comparison to pre-stored prototypes during recognition, while the latter often uses probabilistic graphical models such as HMMs (hidden Markov models) [16] and CRFs (conditional random fields) [15] to directly model temporal signals of actions.

In response to practical difficulties of feature tracking, features based on silhouettes or interest points are more popular in the recent literature. This work uses information derived from space-time silhouettes to characterize human actions. Considering that state-space models generally involve complex mathematical and statistical computation (*e.g.*, parameter learning on large training data), this work prefers the template-matching strategy, which is simpler, as well as enabling the use of any ‘static’ classifiers such as the commonly used k -nearest neighbor (k NN) and the more sophisticated support vector machines (SVM). In summary, the main contribution of this paper is to propose the use of the quantized vocabulary of averaged silhouettes to encode human actions. Although the idea behind our method is simple to understand (and easy to implement), its performance in recognizing actions is surprisingly satisfactory, compared with the state-of-the-art methods.

2. Our Method

2.1 Silhouette extraction and normalization

The different deformations of the human silhouette over time can be used as discriminating features to describe actions. Silhouettes are also easy to extract and are insensitive to the foreground texture and color. As a preprocessing step, we first convert motion information in raw action videos to an associated sequence of

silhouettes, which implicitly reflect motion dynamics. Given an action video consisting of T image frames $\mathbf{V} = \{\mathbf{I}(x, y, t)\}_{t=1}^T$, we assume that the associated sequence of silhouettes $\{\mathbf{S}(x, y, t)\}_{t=1}^T$ can be obtained either by motion detection techniques such as background subtraction and temporal differencing or by a contour tracker [4]. The size and position of the silhouette region vary with the distance of the human from the camera, the human size and the action being performed. In our view, the speed of global translation and the scale difference are less informative for action recognition than the shape and speed of the limbs relative to the torso. Thus silhouettes are centered and normalized on the basis of preserving the spatial aspect ratio of the silhouettes. The resulting normalized silhouettes $\{\hat{\mathbf{S}}(x, y, t)\}_{t=1}^T$ contain as much foreground as possible, do not distort the moving shape and have the same dimensions of $n_1 \times n_2$ (we set $n_1 = 64$ and $n_2 = 48$ in experiments). Figure 1(a) gives an example of normalized silhouette sequences, in which we display every other 4 frames.

2.2 Averaged silhouettes

To capture local temporal characteristics of space-time silhouettes induced by consecutive changes of silhouettes (as well as global shape information of each silhouette), we wish to divide any sequence of normalized silhouettes into subsequences of a window size w with an overlap l between any two consecutive subsequences. In our experiments, we generally set $l = w/2$, like [2, 5, 7]. Let the starting frame index be s , the spatiotemporally windowed subsequence can be denoted as

$$\hat{\mathbf{S}}_s = \{\hat{\mathbf{S}}(x, y, s), \dots, \hat{\mathbf{S}}(x, y, s + w - 1)\}.$$

For each subsequence, we average the silhouettes to arrive at a set of average silhouettes by

$$\mathbf{AS}_s = \frac{1}{w} \sum_{i=0}^{w-1} \hat{\mathbf{S}}(x, y, s + i).$$

Figure 1(b) shows an example of average silhouettes. Such a representation not only enables the comparison of two subsequences using any dissimilarity metric such as Euclidean distance, but implicitly captures the shape of the body part and, to a lesser extent, the temporal dynamics of action segments (the time spent at each stance shows up indirectly as intensity in the image).

2.3 Averaged silhouette clusters

We wish to create a library of averaged silhouette clusters (ASC) to quantize each action sequence. This

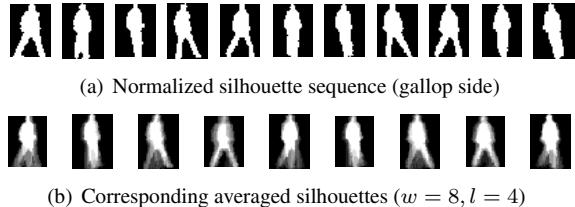


Figure 1. Averaged silhouettes.

vocabulary is realized by clustering a large number of averaged silhouettes extracted from the training videos from all classes of actions. To this end, we first convert each of m averaged silhouettes into a $d = n_1 \times n_2$ dimensional vector in a row-scan manner, and then group them using the k -means clustering algorithm. Note that the number of clusters must be small to avoid learning and recognition in a high dimensional space, while the set of ASCs must contain enough elements to account for variations within and between classes. The resulting clusters of averaged silhouettes tend to be perceptually meaningful, and may considerably discriminate body shapes and action segments.

2.4 Action characterization and classification

After obtaining k averaged silhouette clusters, each averaged silhouette from the original input videos is assigned a type by mapping it to the closest cluster, at which point the averaged silhouettes themselves are discarded and only their types are kept. We use a histogram of the averaged silhouette types as the action descriptor. That is, for each action sequence, we count the frequencies of each ASC occurring in the sequence. The resulting normalized action histogram will be a k -dimensional vector. Figure 2 shows examples which demonstrate that only actions from the same category have a similar distribution and each category has a few dominating averaged silhouette clusters.

Assume that we have n reference action sequences with the corresponding class labels $c_j|_{j=1}^n$ and the histogram-based representations $\{\mathbf{h}_r^j\}_{j=1}^n$. Given a test sequence we can obtain its quantized histogram representation \mathbf{h}_t , the distance between the test sequence and any reference sequences may be calculated by using the χ^2 distance, *i.e.*,

$$d_{\chi^2}(\mathbf{h}_r, \mathbf{h}_t) = \frac{1}{2} \sum_i \frac{(\mathbf{h}_r(i) - \mathbf{h}_t(i))^2}{\mathbf{h}_r(i) + \mathbf{h}_t(i) + \epsilon}$$

where the introduction of a non-zero constant ϵ is just to avoid “divided by zero” in practice. The class of the test sequence will be determined as the class of the reference sequence with the minimum distance by

$$c_{j'} = \arg \min_{j=1 \sim n} d_{\chi^2}(\mathbf{h}_r^j, \mathbf{h}_t).$$

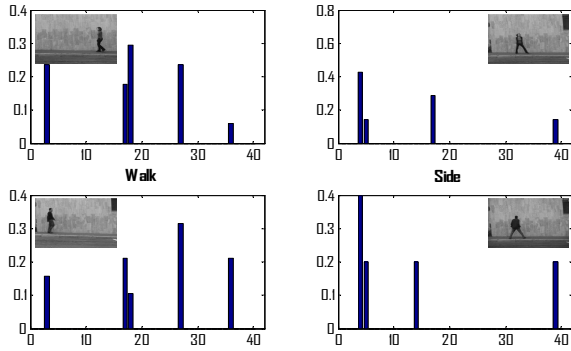


Figure 2. Examples of action histograms.



Figure 3. Examples of 10 actions.

3. Experimental Results

Evaluation data: Our approach was tested on the widely used Weizmann data set, which is appreciably sized in terms of the number of people, actions and video sequences, and is realistic and challenging, including inter-person variations due to different physical bodies and motion styles. The early Weizmann data set used in [2] contains 81 low-resolution videos (180×144 , 25fps), from 9 different people, each performing 9 actions, *i.e.*, run, walk, jumping-jack (or jack), jump-forward-on-two-legs (or jump), jump-in-place-on-two-legs (or pjump), gallop side ways (or side), wave-two-hands (or wave2), wave-one-hand (or wave1), and bend. The extended data set in [5] includes the “skip” action from these 9 people, 90 videos in total. See Figure 3 for examples. Silhouette masks and original image sequences of 93 videos are provided, including three additional videos, *i.e.*, run, skip and walk.

Results: We directly used silhouette masks provided for our experiments. As mentioned in [5], these silhouettes contained “leaks” and “intrusions” due to imperfect subtraction, shadows, and color similarities with the background. Although they are not very perfect, they suffice for our method. We use the leave-one-out cross validation scheme and the nearest neighbor classifier in our experiments. That is, each time we leave one sequence out for testing and the remaining 92 sequences for training. This process is repeated 93 times.

In addition to $w = 10, l = 5$ used in [2, 7] and $w = 8, l = 4$ in [5], we tried $w = 6$ and $w = 12$. We set the range of k according to the approximate num-

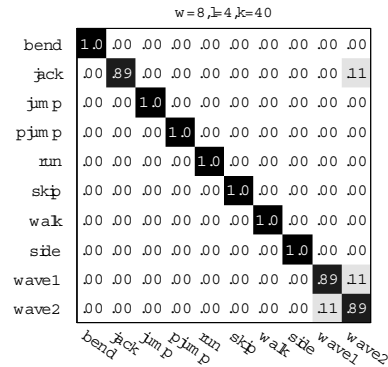


Figure 4. Confusion matrix.

ber of averaged silhouettes available in the training data (say \tilde{m}). Since the performance of the k -means clustering is subject to random initializations of cluster centers, for each parameter setting we performed 10 trials, and reported the best and average recognition rates. The main results are summarized in Table 1, from which we can see that our method can obtain satisfactory accuracies. In addition, the combination of $w = 8, l = 4$ and $k = 40$ performs best overall. We also showed the confusion matrix of action classification in the best case in Figure 4, in which only one ‘wave 1’ and ‘wave 2’ were confused with each other and one ‘jack’ was wrongly classified to ‘wave2’.

Comparison: We compared some recent algorithms, especially those using the same data set. These algorithms used different visual cues to extract features for recognition, *e.g.*, silhouettes [15, 9], shapes (or contours) [5, 7, 16, 6], trajectories [1], optical flow [3, 8], gradients [13, 10], and interest points [11, 12]. It should be noted that although the results are all on the Weizmann data set, these methods are slightly different with regard to the feature extraction and representation schemes, classification methods as well as training and testing data settings.

We directly listed the ‘best’ results reported by these algorithms in Table 2. It can be seen that on the 93-sequence data, our method’s performance is very close to those of [3] and [14] which used optical flow and/or silhouettes for extracting features and more sophisticated AdaBoost (or 1NN-M) for classification, and is better than all others. On the 90-sequence data set, the method of [5] obtained fully correct recognition. Since we do not know which 9 sequences were used among each of actions of run, walk and skip, we cannot do exact experiments for comparison. Jhuang *et al.* [8] obtained 98.8% accuracy, but the experiment was performed on the 81 sequence data without the skip action. In summary, our method is competitive to the state-of-the-art methods. The advantages of our method are the simplicity of feature extraction and characterization,

Table 1. Recognition rates (%) over 10 trials of k -means

(w, l) [\bar{m}]	Average					Best				
	$k = 20$	$k = 30$	$k = 40$	$k = 50$	$k = 60$	$k = 20$	$k = 30$	$k = 40$	$k = 50$	$k = 60$
(6, 3) [1758]	89.8	89.9	91.9	91.6	89.1	94.6	94.6	95.7	94.6	94.6
(8, 4) [1280]	89.8	90.3	93.0	90.3	89.3	93.6	93.6	96.8	95.7	93.6
(10, 5) [995]	88.2	87.7	90.3	88.7	87.0	91.4	92.5	94.6	92.5	91.4
(12, 6) [672]	87.6	90.1	86.8	85.1	83.2	92.5	96.8	90.3	89.3	86.0

Table 2. Performance comparison of different approaches on the Weizmann data set.

Methods	# Seq.	Accuracy(%)	Classifier	Basic visual cue(s)
Liu <i>et al.</i> 2008 [11]	81	86.8	5NN	Interest points and shapes
Ali <i>et al.</i> 2007 [1]	81	92.6	5NN	Trajectories
Jhuang <i>et al.</i> 2007 [8]	81	98.8	Linear SVM	Shapes and motions
Niebles and Li 2007 [12]	83	72.8	Linear SVM	Shapes and interest points
Gorelick <i>et al.</i> 2007 [5]	90	100	1NN	Shapes
Wang and Suter 2007 [15]	90	97.8	Factorial CRF	Silhouettes
Thureau and Hlavac 2008 [13]	90	94.4	1NN	Oriented gradients
Hsiao <i>et al.</i> 2008 [7]	90	96.7	1NN	Shapes
Grundmann <i>et al.</i> 2008 [6]	90	94.6	1NN	Shapes
Yu and Aggarwal 2009 [16]	93	93.6	HMM	Shapes
Kläser <i>et al.</i> 2008 [10]	93	84.3	Nonlinear SVM	Oriented 3D gradients
Jia and Yeung 2008 [9]	93	90.9	kNN	Silhouettes
Fathi and Mori 2008 [3]	93	100	AdaBoost	Optical flow
Tran and Sorokin 2008 [14]	93	100	1NN-M	Optical flow and silhouettes
Our approach	93	96.8	1NN	Silhouettes

avoiding explicit tracking and complex computation of space-time interest points or optical flow, and computationally expensive state-space models.

4. Summary

This paper has proposed a simple yet powerful solution to human action recognition. The main idea is the use of a quantized vocabulary of averaged silhouettes for encoding actions. Our approach has shown a clear advantage over other start-of-the-art methods. Future works include the use of more sophisticated clustering algorithms, classifiers and multiple visual cues, as well as validating the algorithm on a larger data set that exhibits more variations.

References

- [1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *ICCV*, 2007.
- [2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.
- [3] A. Fathi and G. Mori. Action recognizing by learning mid-level motion features. In *CVPR*, 2008.
- [4] D. M. Gavrilu. The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
- [5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *TPAMI*, 29(12):2247–2253, 2007.
- [6] M. Grundmann, F. Meier, and I. Essa. 3D shape context and distance transform for action recognition. In *ICPR*, 2008.
- [7] P. C. Hsiao, C. S. Chen, and L. W. Chang. Human action recognition using temporal-state shape contexts. In *ICPR*, 2008.
- [8] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.
- [9] K. Jia and D. Y. Yeung. Human action recognition using local spatio-temporal discriminant embedding. In *CVPR*, 2008.
- [10] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.
- [11] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *CVPR*, 2008.
- [12] J. Niebles and F. Li. A hierarchical model of shape and appearance for human action classification. In *CVPR*, 2007.
- [13] C. Thureau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.
- [14] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *ECCV*, 2008.
- [15] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *CVPR*, 2007.
- [16] E. Yu and J. Aggarwal. Human action recognition with extremities as semantic posture representation. In *CVPR*, 2009.