Informative Shape Representations for Human Action Recognition*

Liang Wang and David Suter
ARC Centre for Perceptive and Intelligent Machines in Complex Environments
Monash University, Clayton, VIC, 3800, Australia
{liang.wang, d.suter}@eng.monash.edu.au

Abstract

Shape and kinematics are two important cues in human movement analysis. Due to real difficulties in extracting kinematics from videos accurately, this paper proposes to address the problem of human action recognition by spatiotemporal shape analysis. Without explicit feature tracking and complex probabilistic modeling of human movements, we directly convert an associated sequence of human silhouettes derived from videos into two types of computationally efficient representations, i.e., average motion energy and mean motion shape, to characterize actions. Supervised pattern classification techniques using various distance measures are used for recognition. The encouraging experimental results are obtained on a recent dataset including 10 different actions from 9 subjects.

1. Introduction

Visual analysis of human movements [1] concerns the detection, tracking and recognition of people, and more generally, understanding of human activities. In particular, human action recognition has a wide range of promising applications such as surveillance, perceptual interface, analysis of sport events, etc. In the recent literature, there has been considerable work on human action recognition [3-15], which generally fall under two major categories. One is template matching based approaches, e.g. Bobick and Davis [4] proposed a view-based approach to the representation and recognition of aerobics actions using temporal templates, and the other is state-space methods, e.g., Yamato *et al*. [8] combined the mesh features of human blobs with HMMs to identify tennis behaviors.

Past work usually involves the computation of optical flow [3,6,7] and intensity-based features [9], key-frame detection [5], feature tracking [12], etc. Schuldt *et al*. [9] constructed video representations in terms of local space-time features for action recognition. Efros *et al*. [7] introduced a spatiotemporal descriptor of optical flow measurements for recognizing actions. In [12], trajectories of various parameters of a kinematic body model were used for gait classification. However, these studies based on tracking, intensity-based features and the computation of local space-time gradients might be unreliable with respect to low-quality videos, motion discontinuities, singularities, changes of appearances, self-occlusions, etc.

Human actions are essentially spatiotemporal variations of human silhouettes which encode spatial information of human body poses and dynamic information of body motions. The methods that use motion features that could be directly derived from the space-time silhouettes are becoming popular [4,10,11,14], e.g., Blank *et al*. [11] used the properties of solution of the Poisson equation of the space-time silhouette volume to extract features for action recognition. In this paper, to characterize an action, we directly convert an associated sequence of human motion silhouettes derived from the video into two kinds of compact and informative representations which implicitly capture the global motion properties of the human body and motion patterns of local body parts. The advantages of such representations are low computational complexity and simple implementation. Also, explicit feature tracking and complex probabilistic modeling of human motions are not required. Experimental results on a relatively larger action dataset provide the encouraging performance.

2. Feature extraction and representation

Our basic assumption is that a sequence of regions of interest (ROI) containing a moving human can be obtained. Then, two computationally efficient representations, i.e., average motion energy (AME) and mean motion shape (MMS), are extracted. They indirectly encode the motion structure and characteristics of an action, and save both storage space and computation complexity.

2.1. Average motion energy

Given a sequence of binary silhouette images involving the moving human $B_t(x, y)$, the AME is defined by

$$A = \frac{1}{\tau}\sum_{t=1}^{\tau} B_t(x, y) \quad (1)$$

where τ is the duration (i.e., the number of frames) of a complete action. For periodic actions, it may be simply

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

chosen as a cycle(s). There have been many studies for periodicity detection of motions, e.g., a robust method based on image self-similarity [16]. Unlike some other representations that consider motion as a sequence of static poses, the AME represents a human motion sequence in a gray-level image while preserving some temporal information, and has been successfully used in gait-based human identification [15] (called gait energy image (GEI) there) and repetitive activity recognition (walking and running with global horizontal motions) in thermal imaginary [14]. It is believable that such a representation should not be limited only to repetitive motions. Thus, we extend its use to represent any complex movements, e.g., periodic or non-periodic actions, with or without global motion in both horizontal and vertical directions.

Some silhouette images in a jumping-jack action are shown in Figure 1, and the right most is the associated AME, from which we can see that it reflects major shape of silhouettes and their changes over the motion duration. A pixel with higher intensity means that motion occurs more frequently at this position.

Figure 1. An example of AME representation

2.2. Mean motion shape

The MMS is proposed based on shapes, not silhouettes, in a similar manner to the AME. The boundary or shape can be easily obtained from the single-connectivity binary silhouette using a border following algorithm. First, we compute its shape centroid (x_c, y_c). Let the centroid be the origin of 2D shape space. We can represent each shape as a set of boundary points in a common complex coordinate, i.e., a vector with k complex numbers, $\boldsymbol{z}=[z_1, z_2, \ldots, z_k]^T$. Each action is accordingly converted into a sequence of such 2D shape configurations.

To compare a set of static shapes in movement patterns with robustness to position, scale and slight rotation changes, a mathematically elegant way, Procrustes shape analysis, is used (for a good summary, the readers may refer to [2]). The centered configuration of a shape is defined as $\boldsymbol{s}=[u_1, u_2, \ldots, u_k]^T, u_i = z_i - \bar{z}$. Given a set of n motion shapes, their mean shape can be obtained by finding $\boldsymbol{s}$ to minimize the objective function

$$\min_{\alpha_j,\beta_j} \sum_{j=1}^{n} \left\| \boldsymbol{s} - \alpha_j \boldsymbol{1}_k - \beta_j \boldsymbol{s}_j \right\|^2 \qquad (2)$$

where $\alpha_j \boldsymbol{1}_k$ translates $\boldsymbol{s}_j$, and $\beta_j = |\beta_j| e^{i\angle\beta_j}$ scales and rotates $\boldsymbol{s}_j$. To find $\boldsymbol{s}$, the following matrix is computed.

$$\boldsymbol{E}_s = \sum_{j=1}^{n} (\boldsymbol{s}_j \boldsymbol{s}_j^T) / (\boldsymbol{s}_j^T \boldsymbol{s}_j) \qquad (3)$$

The Procrustes mean shape is the eigenvector that corresponds to the greatest eigenvalue of $\boldsymbol{E}_s$ [2]. Figure 2 shows some 2D shapes in a running action, and the right most is the associated MMS, which reflects major shape of the human body and their deformations. Positions where the shape varies more compared with general body shape indicate regions in which the associated limbs move more.

Figure 2. An example of MMS representation

3. Classifier

Since our interest is to evaluate the real discriminatory powers of the used feature representations, only three simple classification methods, namely the nearest neighbor classifier (NN), the k-nearest-neighbor classifier (kNN), and the nearest neighbor classifier with class exemplar (ENN), are adopted here.

To measure similarity between the test action T and the reference action R, various distance metrics are used. For the AME, the summation of absolute difference (SAD) and the simplified Mahalanobis distance are used.

$$d_{AME} = \sum_{x,y} \left| A^T_{x,y} - A^R_{x,y} \right| \quad \text{or} \quad d_{AME} = \sum_{x,y} \frac{\left(A^T_{x,y} - A^R_{x,y}\right)^2}{\sigma^2_{x,y}} \qquad (4)$$

For the MMS, the full Procrustes shape distance is used.

$$d_{MMS} = 1 - \frac{\left| S_T^* S_R \right|^2}{\left\| S_T \right\|^2 \left\| S_R \right\|^2} \qquad (5)$$

The smaller the above distance measures are, the more similar the two actions are.

4. Experiments

4.1. Dataset

For evaluating the proposed algorithm, this paper uses a recent database reported in [11], which is relatively larger in terms of the number of subjects and actions. It includes 81 low-resolution videos (180-by-144, 25fps) from 9 different people, each performing 9 natural actions (periodic and non-periodic actions, and stationary and non-stationary motions along both horizontal and vertical directions), i.e., bending (bend), jumping jack (jack), jumping-forward-on-two-legs (jump), running (run), jumping-in-place-on-two-legs (pjump), walking (walk), galloping-sideways (side), waving-one-hand (wave1), and waving-two-hands (wave2). Together with a lately added action of skipping (skip), the dataset in our experiments in total includes 10 actions and 90 videos. Sample images of each action are shown in Figure 3, from which we can see that many of actions are similar in the senses that the limbs have similar motion paths, and this high degree of similarity among actions makes discrimination more challenging. Also, each action in this dataset is performed

by different people with different physical characteristics and motion styles, thus providing more realistic data for the test of the versatility of the proposed method.

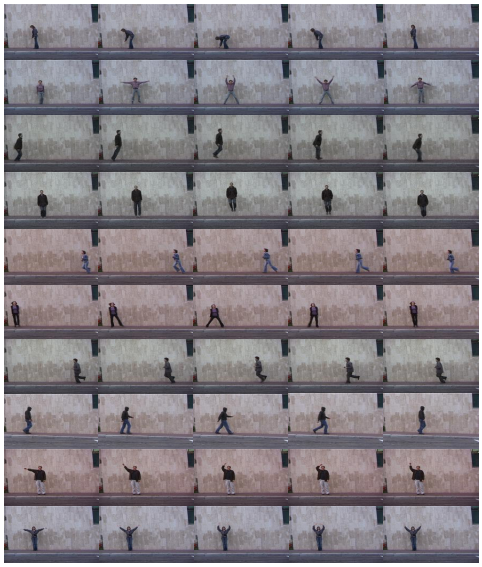


Figure 3. Sample images in the action database. From top to bottom: bend, jack, jump, pjump, run, side, skip, walk, wave1, and wave2, respectively

4.2. Processing

We directly use the masks from [11] for the subsequent processing. These foreground silhouettes contain leaks and intrusions due to imperfect subtraction, shadows and color similarities with the background. To emphasize motion of parts relative to the torso, the translation of the center of mass for actions with global motions is compensated by fitting a second order polynomial to the frame centers of mass [11].

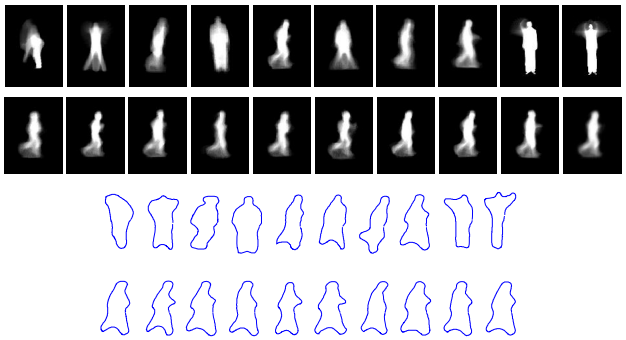
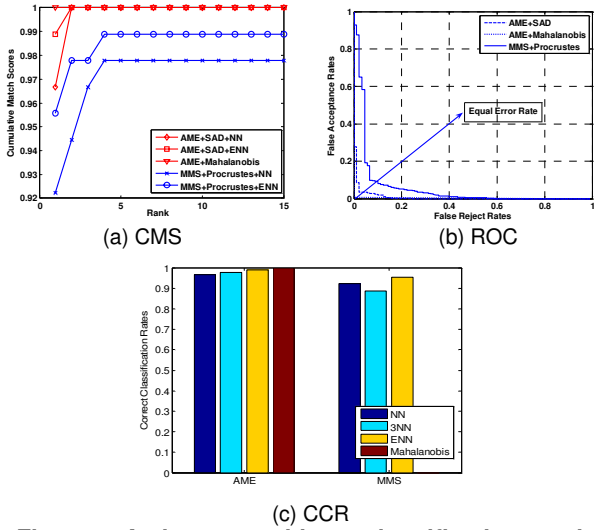


Figure 4. First and third rows: AMEs and MMSs of 10 different actions from a same subject, respectively; second and fourth rows: AMEs of a running action and MMSs of a walking action from 9 different subjects and their exemplars (the right most), respectively

Each silhouette sequence is then converted into both AME and MMS representations in the manner described in Section 2. Further, we use the class average of the AME and the MMS derived from multiple sequences with the same action as the exemplar for the corresponding action class. Figure 4 shows some plots of AME, MMS and their exemplars, from which it can be seen that these two representations have significant discriminating powers.

4.3. Results

For a small number of examples, the leave-one-out cross-validation rule is adopted to compute an unbiased estimate of the true recognition rate. A useful classification performance measure, CMS (cumulative match score), is used for the evaluation of action classification. It is defined as the cumulative probability $p_{(k)}$ that the real class of a test is among its top k matches. For completeness, we also estimate FAR (false acceptance rate) and FRR (false reject rate) via the leave-one-out rule. Note that, each time there is one genuine attempt and 9 imposters since the left-out sample is known to belong to one of the 10 action classes. Figure 5 shows the CMS up to 15 and ROC, from which we can see that the AME plus Mahalanobis performs best for both identification and verification.



(a) CMS (b) ROC

(c) CCR

Figure 5. Action recognition and verification results

It should be noted that the correct classification rate (*CCR*) is equivalent to $p_{(1)}$ (i.e., rank=1). For clarity, the CCRs using different representations and classifiers are summarized in Figure 5(c), from which it can be seen that: 1) the AME representation generally outperforms the MMS. This is probably because the AME preserves considerable temporal information of motion. It is also less sensitive to the segmentation noises in individual frame. However the MMS representation is somewhat limited in these two aspects. 2) The ENN classifier generally performs better than the NN classifier. For each action, there are slightly changes between the action sequences performed by different people with different physical structures and motion forms. The average of multiple same-action sequences might provide a more standard pattern for that

action than using only a single and random action sample.

To examine and analyze which action sequences are incorrectly classified, we specifically show two confusion matrixes with respect to the NN classifier in Figure 6. For the AME plus the NN, only three sequences are incorrectly classified (i.e., one jump, one skip and one wave1 are classified as skip, jump and wave2, respectively). In fact, jump and skip, wave1 and wave2 are very similar in moving paths of local limbs. For the MMS plus the NN, seven sequences are wrongly classified (i.e., one jump, one side, one walk, one wave1, and one wave2, are respectively classified as pjump, walk, run, wave2, and jack; and two pjumps are classified as jump and side). In addition to the high similarities between motions, poor foreground segmentation in these sequences contributes to the confusion.

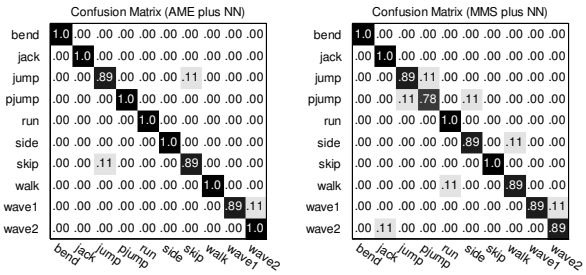


Figure 6. Confusion matrixes of action recognition

4.4. Discussion

Compared with most past work, the proposed method has several advantages: 1) it is easier to comprehend and implement, without the requirement of explicit feature tracking and complex probabilistic modeling; 2) as a silhouette-based method, it is insensitive to color and texture of cloth. It naturally avoids some problems in previous methods such as unreliable feature tracking due to self-occlusion and appearance changes, and complex optical flow computation; and 3) it obtains good results on a larger database. Our results are comparable to those of Blank *et al*. [11] (a recognition rate of almost 100% on the same database of 81 sequences without skipping is reported there), and our feature selection and extraction are simpler than theirs.

Much work remains open: 1) although the results are encouraging, evaluations on a larger and realistic database need to be investigated in order to be more conclusive; 2) the work by Veeraraghavan *et al*. [10] has showed that for the task of activity recognition, models that encode both shape and kinematics are required. An interesting attempt will be to exploit the kinematics in human movements using time-sequential networks, though the kinematics is not explicitly modelled or captured; 3) both shape and kinematics information play important roles for motion analysis. Fusion of these two cues is thus preferable for improving the accuracy and reliability; and 4) more robust articulated shape matching and foreground segmentation are also parts of future work.

5. Conclusions

Compared with other widely studied topics such as human detection, tracking and recognition, human activity understanding is in its infancy. This paper has addressed the problem of human action recognition using two kinds of informative spatiotemporal representations derived from motion silhouettes or shapes. Extensive experimental results have demonstrated their powerful capabilities in human action recognition.

References

[1] D. Gavrila. The visual analysis of human movement: a survey. *CVIU*, (1999) 73(1): 82–98

[2] J. Boyd. Video phase-locked loops in gait recognition. *CVPR* (2001): 696-703

[3] M. Black. Explaining optical flow events with parameterized spatiotemporal models. *CVPR* (1999): 1326-1332

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

[5] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. *Workshop on Models versus Exemplars in Computer Vision* (2001)

[6] O. Boiman and M. Irani, Detecting irregularities in images and in videos, I: 462-469, *ICCV* (2005)

[7] A. Efros, *et al*. Recognizing action at a distance. *ICCV* (2003): 726-733

[8] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time sequential images using hidden markov model. *CVPR*, (1992): 379-385

[9] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. *ICPR*, (2004) 3: 32–36

[10] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa. Role of shape and kinematics in human movement analysis. *CVPR* (2004): 730-737

[11] M. Blank, *et al*. Action as space-time shapes. *ICCV* (2005): 1395-1402

[12] A. Bissacco, *et al*. Recognition of human gaits. *CVPR*, (2001) 2: 52-57

[13] A. Yilmaz and M. Shah. Action sketch: a novel action representation. *CVPR* (2005): 984-989

[14] J. Han and B. Bhanu. Human activity recognition in thermal infrared imagery. *Workshop on Object Tracking and Classification Beyond the Visible Spectrum* (2005)

[15] J. Han and B. Bhanu. Statistical feature fusion for gait-based human recognition. *CVPR*, (2004) 2: 842–847

[16] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *PAMI*, 22(8) (2000): 781-796

* This work is supported by ARC Centre for Perceptive and Intelligent Machines in Complex Environments, Monash University, Australia.