DOMINANT SETS-BASED ACTION RECOGNITION USING IMAGE SEQUENCE MATCHING

Qingdi Wei, Weiming Hu, Xiaoqin Zhang, Guan Luo

National Laboratory of Pattern Recognition
Institute of Automation, CAS, Beijing, China
{qdwei,wmhu,xqzhang,gluo}@nlpr.ia.ac.cn

ABSTRACT

Action recognition is one of the most active research fields in computer vision. In this paper, we propose a novel method for classifying human actions in a series of image sequences containing certain actions. Human action in image sequences can be recognized by a time-varying contour of human body. We first extract shape context of each contour to form the feature space. Then the dominant sets approach is used for feature clustering and classification to obtain the labeled sequences. Finally, we use a smoothing algorithm upon the labeled sequences to recognize human actions. The proposed dominant sets-based approach has been tested in comparison to three classical methods: K-means, mean shift, and Fuzzy-Cmean. Experimental results demonstrate that the dominant sets-based approach achieves the best recognition performance. Moreover, our method is robust to non-rigid deformations, significant scale changes, high action irregularities, and low quality video.

Index Terms— Image motion analysis

1. INTRODUCTION

Action recognition has been received more and more attentions due to its crucial values in video surveillance and monitoring, human-computer interactions, video indexing and browsing, etc. Despite the increasing amount of work done in this field in recent years, action recognition remains a challenging task for the following reasons: (i) it is difficult to find a general descriptor for human action, because its non-rigid and high degree of freedom essences. (ii) the length of action period is variational, which poses the problem of action segmentation. (iii) nuisance factors, such as self-occlusion, low quality video and irregularity of camera parameters also bring extra difficulties.

In this paper, we propose a novel approach which does not need action segmentation as traditional methods. In our approach, action recognition is to extract Shape Context[1, 2, 3] from human contours frame by frame first, and then cluster action features to build a database by Dominant Sets[4, 5], which is a novel approach based on graph theory, finally classify the test action. Thus, we can achieve a high recognition rate without having to consider the length of action period.

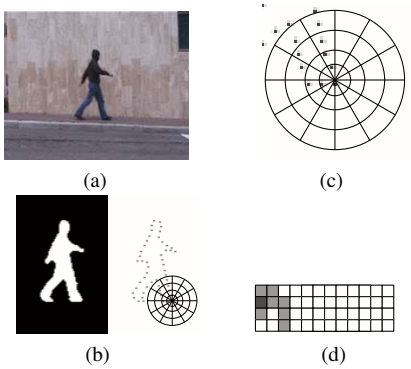

Fig. 1. (a) original image (b) background subtraction and contour (c) shape context of a point of the contour (d) each shape context is a log-polar histogram of the coordinates of the rest of the point set measured using the reference point as the origin(Dark=large value).

The rest of the paper is organized as follows. Related works are discussed in Section 2. section 3 will deal with the major problem. Action feature and action recognition are respectively presented in Section 3.1 and Section 3.2. Experimental results are shown in Section 4, and Section 5 is de-voted to conclusion.

2. RELATED WORK

Roughly speaking, previous work on action recognition can be classified into two categories. One is based on extracting 2D features to describe human action, e.g. contour, color, lines etc. Gorelick et al.[6] suggested using Poisson equation to represent each contour. Liu and Ahuja[7] proposed an alternative way to use of Fourier descriptors. Kale et al.[8] utilized a vector of width to reach the same objective. All researches in the field of 2D features have demonstrated that human contours contain rich information about the shape of human body, leading to some elegant work that human action can be readily recognized based on information extracted from a contour sequences of human actions.

The second category of approaches use 3D features to estimate the human gesture. Irani and his colleagues[9, 10] analyzed actions through treating an video sequence as a space-time intensity volume. Human actions were also represented by a 3D spatio-temporal surface, as in[11, 12, 13].

Most approaches for action recognition mentioned above require solving the problem of varying length of action periods, which is usually not a trivial task. When comparing test sequences with sample sequences, the recognition would be severely affected if they are not well aligned. Comparing to the previous work, our method avoids this problem to reach a higher recognition rate.

3. RECOGNITION METHOD

We are interested in human action recognition through a series of image sequences containing certain actions. $S = \{s_t\}_{t=1}^{T}$ is denoted as the input sequence, where each s_t denote an image at time t. The recognition problem is usually formed as a classification problem. Specifically, our purpose is to find a classifier $f : f(S) = c$ that classifies a given sequence as action $c \in C = \{1, \cdots, n_c\}$, where C is a set of actions that we are interested.

Our algorithm consists of two stages which are training phase and testing phase. Training Phase: 1: Preprocess video sequences to obtain contours of moving region in every frame. 2: Extract shape context from each contour to build a feature sequence. 3: Cluster features sequence to get image classes, which form a sample database. 4: The priori probability of each image class is estimated from the frequency of action classes. Testing Phase: 1: Preprocess test video sequences to obtain contours of moving region in every frame. 2: Extract Shape Context from each contour to get a feature sequence. 3: Classify features in the sequence to get an image class sequence, and then a probability sequence. Here, coarse action recognition is completed. 4: Fine recognition is completed by using a smooth algorithm to reduce noise.

3.1. Action Feature

The contour of movement is obtained during the preprocessing procedure. This objective could be reached by many approaches, e.g. background subtraction, border following and optical flow.

The human contour is described by a feature vector extracted by a shape descriptor. Shape context of a point is a matrix sc_l, which describes points distributed around, as shown in Fig.1. Then the contour is represented by a discrete set $SC = \{sc_l\}_{l=1}^{L}$,containing L points sampled from the external contours, and the image sequence is represented by $ims = \{SC_i\}_{i=1}^{n}$.

Algorithm 1: Dominant-Set Clustering
Input : Affinity matrix for kth iteration A^k
1. If A^k is empty, return $NULL$
2. Calculate the local solution of (1) by (2): u^k and $f(u^k)$
3.Get the dominant set:$S^k = C_{u^k}$
4. Split out S^k from the present graph and get a smaller graph with new affinity matrix A^{k+1}
Output: $A^{k+1}, S^k, u^k, f(u^k)$

Algorithm 2: Dominant-Set Fast Assignment Algorithm
Input :Affinity vector $a \in \mathbb{R}^{n\times 1}, \cup_{k=1}^{K}\{S^k, u^k, f(u^k)\}$
1. $\frac{w_{S^k\cup\{p^{new}\}}(p^{new})}{W(S^k\cup\{p^{new}\})} = \frac{|S^k|-1}{|S^k|+1}\left(\frac{a^T u^k}{f(u^k)} - 1\right)$ for all $k \in \{1, \cdots, K\}$
2. if $w_{S^{k^*}\cup\{p^{new}\}}(p^{new}) > 0$, then assign p^{new} to cluster S^{k^*}
Output: k^*

3.2. Clustering

In the training section, shape context features are clustered by Dominant sets in each image, because of its high purity results.

Dominant set, proposed by Pavan and Pelillo[4], is a novel graph-theoretic approach for clustering and segmentation. It is proved that finding a "dominant set" is equivalent to solve a quadratic program:

$$\begin{array}{ll} \text{maximize} & f(u) = \frac{1}{2}u^T A u \\ \text{subject to} & u \in \Delta \end{array} \tag{1}$$

where

$$\Delta = \{u \in \mathbb{R}^n : u \geq 0 \text{and} \sum_{i=1}^{n} u_i = 1\}$$

and A is the symmetric affinity matrix. If there exists u^*, a strict local solution of the program(1), C_{u^*} is equivalent to a dominant set of the graph, where $C_u = \{i : u_i > 0\}$. *Replicator equation* can be used to solve program(1):

$$u_i(t+1) = u_i(t)\frac{(Au(t))_i}{u(t)^T Au(t)} \tag{2}$$

The dominant sets clustering algorithm is shown in Algorithm 2, illustrating the kth iteration in clustering. A dominant set is split out from the current graph in each iteration, until that the rest of graph can not generate any dominant set. The number of clusters K is therefore automatically determined.

3.3. Classification

Dominant sets can also be used to do classification. In[4], Pavan and Pelillo made an out-of-sample extension for dominant sets clustering. We adopt the idea to classify a new sample p^{new}. Given a new sample p^{new} and of a trained data-base $\bigcup_{k=1}^{K}\{S^k, u^k, f(u^k)\}$, which is the output of Algorithm 1,

we first contrast an affinity vector a which represents the distance between p^{new} and n existing samples. The assignment algorithm is shown in Algorithm 2, where k^* is the cluster to which p^{new} belongs.

3.4. Recognition

In this paper, recognition is realized by a process of classification. Dozens of classes of per-frame images are obtained by the method in Sec. 3.2. Since an image class is rarely comprised of images belonging to the same action class, it is reasonable that the probability transformation T_i can be estimated by the frequencies of each action class appearing in an image class.

$$T_i : image\ class\ i \rightarrow \begin{cases} action\ class\ 1 & p_{i1} \\ \vdots & \vdots \\ action\ class\ n & p_{in} \end{cases} \quad (3)$$

where

$$\sum_{j=1}^{n} p_{ij} = 1$$

Using dominant sets, we classify each image in the test sequence to get an image class sequence. Then the transformation is applied to the image class sequence: *action class sequence* = *T* (*image class sequence*). Action class sequence is not an $n \times 1$ vector, but an $n \times n$ matrix. Each frame in test video sequence corresponds to a $1 \times n$ vector in action class sequence. The coarse action recognition is done by max action class sequence on each frame.

For a better recognition, we prefer the smoothing algorithm to the max process, because the former can reduce noise. In the smooth algorithm, sum up probabilities of action class sequences in a time section t by equation (4), and then take the $max(LAP)$ as the default class in this time section.

$$LAP(p_1, \cdots, p_n) = \sum_{i=1}^{t} (p_{1i}, \cdots, p_{ni}) \quad (4)$$

Verify images one by one, until that the default class could not fit the broken image. Then repeat the process around broken image. Through the smooth algorithm, we can label the local sequence by the most frequent action class. The smooth algorithm is shown in Algorithm 3.

Algorithm 3: Smooth Algorithm
Input: $actionclass\ sequence$
1, Calculate local actionclass probability LAP, from i to $i + 7$, by equation (4)
2, $default = max(LAP)$
3, Compare $actionclass_{i+1}$ to $default$, if equal then $SIGN$ increase 1, else $SIGN$ decrease 1.
4, $realaction_i = default$, keep $SIGN \leqq 5$.
5,repeat 3,4, Until $SIGN < 0$ repeat 1.
Output: $realactionsequence$

4. EXPERIMENTS

4.1. Simulated Videos

In the first experiment, our algorithm is applied to recognize human action on the CMU database . 45 video sequences in 5 action classes with an image size of 320×240 and a frame rate of 15 frames/sec are selected as samples. 30 video sequences are used for training and the rest 15 for testing. The 5 action classes taken in consideration are 'walk', 'run', 'wipe', 'jump', 'box'. To simplify image preprocessing, the video sequences that we selected are simulated videos, in stead of real human videos. The results are evaluated by the recognition rate, which is defined as the percentage of correctly classified actions.

Besides the proposed approach, three classical clustering techniques are tested for comparison: K-means, mean-shift, and Fuzzy-Cmean. In these experiments, we classified the test data by looking for the nearest cluster centers in Euclidean distance. Table 1 shows the recognition rates with and without the smooth algorithm. For most of the actions, the Dominant Sets-based approach outperforms the other three techniques, except for "wiping", for which the K means-based method and the mean shift-based method achieve the similar result with Dominant sets, and "boxing", for which FCM demonstrates higher recognition rate. It is obvious that the three classical techniques cannot distinguish the five actions in the experiment, which means that they can not work well in Euclidean space to distinguish human action. By contrast, the dominant sets-based approach demonstrates its efficiency in the entire experiment. It can classify most images correctly, and the score can be higher after smoothing.

Now the reasons why the experimental results happen are investigated. As we mentioned in Section 3.1, the contour is represented by a discrete set $SC = \{sc_l\}_{l=1}^{L}$, so the SC can be considered as a curve. Hence, to match SC is equal to match curve. The unsatisfying results of the three techniques may be due to the poor performance of Euclidean distance on matching curve. In contrast, Dominant set has quite a good overall performance because this clustering method simultaneously emphasizes on *internal homogeneity* and *external inhomogeneity* [4]. However, we observe that all the four approaches almost do not work in the 'jumping' action. It can be explained by the fact that we have not yet taken into consideration the displacement of contour in the sequence. On this condition, "jumping" is misclassified into the "standing" category, which is the common area of "boxing" and "wiping". Besides, in our experiment, smoothing plays a catalyzing role: when ordinary recognition rate is over a certain threshold, smoothing algorithm can optimize the recognition rate by reducing noise; inversely, when ordinary recognition rate is quite low, smooth algorithm may worsen the performance.

4.2. Real videos

To demonstrate that my approach can also work on real videos, we also conducted experiments on Irani's database[9]. It has 81 low-resolution video sequences showing nine different people, each performing nine natural actions such as "running", "walking", "jumping", "jumping forward on two legs", "jumping in place on two legs", "galloping sideways", "waving two hands", "waving one hand", "bending". In the database, 5 people's actions are used for training and 4 people's actions are testing data. The video sequences have 180×144 pixel resolution and 25 frames per second. Experimental results show that the action recognition rate attains 88.89%, with only one person's actions not totally recognized. The misclassification is mainly due to the deformity of contours issued from simple technique of background subtraction. Although our recognition rate was not extremely high, the advantage of our method resides in the fact that we can avoid considering the length of action period. In addition, Irani's experimental data is single action, while our method deals with action sequence containing multi-actions. Thus, Dominant Sets based approach is demonstrated to be able to work in real videos composed of multi-actions.

4.3. Robustness

In this experiment, we also demonstrate that Dominant Sets-based approach is robust to non-rigid deformities, partial occlusions and other defects in video sequences. In Irani's database [9], there are ten videos of walking in different backgrounds, which are difficult to subtract. Our method only failed in 1 video out of 10. As we have not taken into account the displacement of contour in the sequence, the person swinging hands with great dimension is misclassified into "jacking" category.

5. CONCLUSION

In this paper, we have presented a novel method for action recognition, and it has been proved to be very effective by experimental results. Dominant sets have such a good performance in clustering and classification, even in noisy data, that it can be used in both simulated videos and real videos.

Our approach has several advantages. First, it is robust to the variation of action duration, because our approach focuses on per-frame image. Second, it has the potential to be applied in videos comprising various kinds of actions. Finally,this approach can also be used without great change on general 3D shapes.

At the present, our approach does not consider the displacement of contour in the sequence. So we can not distinguish directed actions. In other words, we can only recognize "waving hands", but we can not discern "waving up" and "waving down" yet. Therefore, in future, we will try to introduce displacement features into our approach, in order to recognize directed actions.

	Kmeans	Mean Shift	FCM	Dominant Sets	Dom-Sets with smooth
Walk	1.92	21.15	7.69	78.85	100.00
Run	1.65	0	0	74.38	100.00
wipe	96.35	95.62	0	97.08	100.00
jump	0	0	0	50.00	67.50
box	6.41	0	99.36	57.05	78.85

Table 1: recognition rates.

6. ACKNOWLEDGMENTS

This work is partly supported by NSFC (Grant No. 60520120099 and 60672040) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

The CMU datebase was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

7. REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha. "Shape Matching and Object Recognition Using Shape Context", *IEEE Trans. PAMI.*, 24(24):509-522, 2002.

[2] H. Ling, and D.W. Jacobs, "Shape Classification Using the Inner-Distance", *IEEE Trans. PAMI.*, 29(2):286-299, 2007.

[3] Yang Wang, Hao Jiang, Mark S. Drew, Ze-Nian Li, and Greg Mori. "Unsupervised Discovery of Action Classes", *CVPR*, Vol-ume 2, Page(s):1654 - 1661, 2006.

[4] M. Pavan, and M. Pelillo. "A new graph-theoretic approach to clustering and segmentation", *CVPR*, volume 1, pages 18-20, 2003.

[5] Wei Hu, and Weiming Hu, "A New Active Learning Framework: Hierarchical Graph-theoretic Clustering", *the 2006 IEEE International Conference on System, Man and Cybernetics SMC*,

[6] Gorelick, L., Galun, M., Sharon, E., Basri, R., and Brandt, "Shape representation and classification using the Poisson equation", *CVPR*, Page(s):II-61 - II-67 Vol.2 2004.

[7] Che-Bin Liu, and Ahuja, N.,"A model for dynamic shape and its applications", *CVPR*, Vol.2, Page(s):II-129 - II-134 Vol.2,2004.

[8] Kale, A., Sundaresan, A., Rajagopalan, A.N., Cuntoor, N.P., Roy-Chowdhury, A.K., Kruger, V., and Chellappa, R., "Identification of humans using gait", *IEEE Trans. IP*, 13(9):1163-1173, 2004.

[9] Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., "Actions as space-time shapes", *ICCV*, Page(s):1395 - 1402 Vol. 2 2005.

[10] Y. Ukrainitz, and M. Irani, "Aligning Sequences and Actions by Maximizing Space-Time Correlations", *ECCV*, May 2006.

[11] Hong Li, Greenspan, M., "Multi-scale Gesture Recognition from Time-Varying Contours", *CVPR*, 2005.

[12] P. Peixoto, J. Goncalves, and H. Araujo. "Real-time gesture recognition system based on contour signatures", *ICPR*, volume 1, pages 447-450, 2002.

[13] L. Wang, T. Tan, H. Ning, and W. Hu. "Silhouette analysis based gait recognition for human identification", *IEEE Trans. PAMI.*, 35(12):1505-1518, 2003.