Efficient Human Action Recognition by Luminance Field Trajectory and Geometry Information

Haomian Zheng[1], Zhu Li[1], and Yun Fu[2]

[1]Dept of Computing, Hong Kong Polytechnic University, Hong Kong, China
[2]BBN Technologies, Cambridge, MA 02138, USA
{cshmzheng,cszli}@comp.polyu.edu.hk, yfu@bbn.com

ABSTRACT

In recent years the video event understanding is an active research topic, with many applications in surveillance, security, and multimedia search and mining. In this paper we focus on the human action recognition problem and propose a new Aligned Projection Distance (APD) approach based on the geometry modeling of video appearance manifold and the human action time series statistics on the geometry information. Experimental results on the KTH database demonstrate the solution to be effective and promising.

Index Terms—Curve-distance approach, video event recognition, machine learning.

1. INTRODUCTION

With the development of computing and communication technologies, the video devices are becoming more and more popular in the market. Vast amount of video data is captured and used for various kinds of applications. To take the advantage of this vast amount of video data, people were trying to apply various methods to some intelligent and meaningful applications, such as surveillance, video analysis for security, and on-line video searching in sports games. Therefore, an efficient solution for video content analysis, *e.g*. detect and recognize the human action, is becoming the critical point in this problem.

There have been a lot of works about this problem in the literature. Traditional approaches focused mainly on object segmentation and motion analysis based solutions [1], which try to detect the object directly by image segmentation and then recognize human actions by learning the object-level spatio-temporal features. However, the main problem for these approaches often suffer from the poor robustness to the appearance variances of human actions such as different subjects, lighting, backgrounds or occlusions.

To solve this problem, spatio-temporal interest points [2] or a Scale Invariant Feature Transform (SIFT) [8] are used for a video representation and then used in object level detection, segmentation and tracking. The human actions, therefore, are recognized from the motion and distribution of interest points via local SVM in [3] or generative graphical models in [5]. The performance of these approaches has been proved to be effective and robust to different variations.

On the other hand, spatio-temporal volume modeling based solutions have also been investigated. Instead of finding pre-selected localized features by interest points, this approach treated the video features in both spatial and temporal domain by applying a 3-D cubic structure [4], which is first detected in single frames and then found in the time direction. The feature selection is integrated into the Tensor CCA learning process in [7], which requires a very complicated pre-processing step in cropping out the given objects.

In a more general video action recognition problem, people focused on designing a system which can automatically recognize several pre-defined human action patterns, even under some variation and occlusions. The most important processing in this approach is to find those useful statistics information which can be utilized to classify those discriminative actions into different patterns. In our previous work [6], a scaled luminance field trajectory modeling and matching solution was proposed and the resulting video clip searching achieves high performance in both precision-recall and response time. The key observation is that the luminance field trajectory of video sequences contains sufficient information for a variety of detection and recognition problems and can be potentially implemented with efficient and robust real-time performance, and a differential luminance field trajectory (DLFT) method was proposed in [11].

In this paper we address the human action recognition problem with a similar scaled appearance modeling approach. Similar human action video clips should span similar trajectories in the scaled appearance space with similar temporal footprints. Therefore, a temporally aligned average projection distance metric is developed for human action recognition.

2. GEOMETRY MODELING OF VIDEO EVENT

In this work, we model video clips of different human actions performed by different subjects under different image formation conditions as manifolds in the scaled appearance space, *i.e*, for an *n*-frame video sequence $\{F_k\}$ with frame size $W \times H$, we treat it as a trajectory passing through points $\{x_k\}$ in $\Re^{W \times H}$, with x_k being the scaled luminance field of the original sequence. The original $W \times H$ dimensional space still contains some unnecessary information. The dimension is also too high to efficiently calculate the distance.

In our method we reduce the number of dimension by two steps. First, a smaller icon is generated by down sampling the video frame to $w \times h$, a significant gain in the complexity issue is obtained while the cost on the performance is not much. After this step, some statistical techniques of dimensionality reduction can be applied, such as Principle Component Analysis (PCA) [10] and Linear Discriminant Analysis (LDA) [9] modeling. For frame F_k, an updated trajectory x_k can be generated by equation (1).

$$x_k = AF_k = [a_1^T, a_1^T, \mathbb{L}\ , a_{wxh}^T]F_k, a_j^T \in R^d \qquad (1)$$

where the subspace projection matrix A, with a size $d \times w \times h$, is obtained from an unsupervised local learning, with the objective of preserving the maximum amount of information, while keeping the number of dimension in an acceptable range as in equation (2).

$$\begin{aligned} a_1^T &= \arg\max_a \operatorname{var}_{F_k}(a^T F_k) \\ a_2^T &= \arg\max_a \operatorname{var}_{F_k}(a^T \tilde{F}_1) \qquad (2) \\ &\cdots\cdots \\ a_k^T &= \arg\max_a \operatorname{var}_{F_k}(a^T \tilde{F}_{k-1}) \end{aligned}$$

where the iteration on subspace bases are based on the data:

$$\tilde{F}_{k-1} = F_k - \sum_{j=1}^{k-1} a_j a_j^T F_k \qquad (3)$$

In our method, the PCA is applied to reduce the huge amount of dimension to 24, and the number of dimension is finally reduced to 5 by carrying on a LDA on the result of PCA. After the dimension reduction, the distance is calculated between these 5-dimensional trajectories, which are quite acceptable in the decision stage. Figure 1 shows an example for the first two dimensions of PCA's result.

In Figure 1, every single point stands for a video frame and each video clip is represented as a trajectory. In this 2-D space, spatial feature is reflected in the different positions. From the figure we can see that video clips containing different actions have different shape of curves, and we can judge by our eyes that some actions are obviously different from each other.

Actually the geometry of these curves already contains sufficient statistics to differentiate the different human actions. Different video events span different luminance field trajectories in $\Re^{w \times h}$, with different action time series statistics on the curve. Video clips of different human actions performed by different subjects under different image formation conditions span a space with complex structure and relationship. Furthermore, the human action has a temporal dimension which is reflected as the footprint on the trajectory. Only a pure vector space trajectory modeling approach cannot capture this temporal behavior since the temporal information is missing.

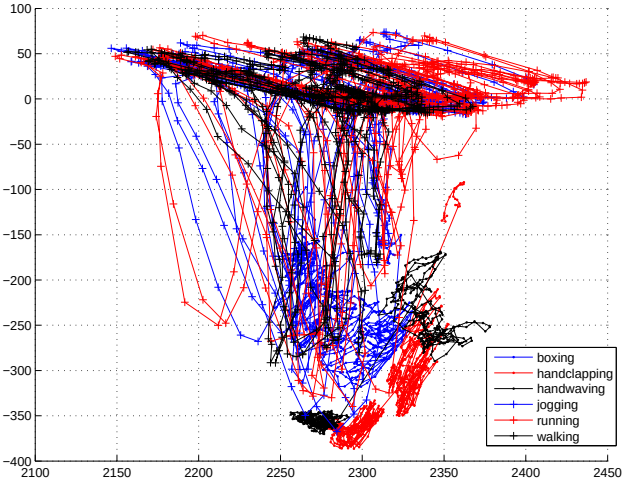

Figure 1: Geometry Information for 6 Human Actions

Motivated by the observations above, we try to model each human action clip by trajectories and footprints in the scaled appearance space, and make human action recognition based on the average projection distance of the video sequences to the trajectory of each action class.

The processing is divided into following steps. First, each video frame is represented as a vector such that it can be projected to a very high dimensional space. Therefore each video clip can be represented as a trajectory which contains sufficient information for further classification. Since not all of the data are contributive to the recognition, the number of dimension is reduced to release the burden on handling the recognition problem. Second, the distances between different curves are calculated, and the matching problem is finally determined based on some relationship between distances, which will be described in Section 3.

3. DISCRIMINATION BASED ON ALIGNED PROJECTION DISTANCE

The classification problem can be divided into two aspects with different properties: temporal property and spatial property.

For the spatial property, which is mainly reflected by the positions in space, the shape of trajectory is considered to be the most important issue on matching and recognition.

Generally speaking, video clips with the same action label will have very similar trajectory and footprint in the geometry information. For each single frame, a similar appearance will become closer points in the trajectory domain.

The temporal property is mainly contained in the footprint of the trajectory. Relationship between any pair of neighbor frames can be found on the trajectory by comparing their positions. Generally, a more obvious change in neighbor frame can be reflected a longer travel in the trajectory domain, and if there is almost no motion, the two neighbor points on the trajectory is very closed to each other.

Based on this observation, we propose the Aligned Projection Distance approach to capture both the spatial and temporal property contained in the trajectory. Therefore, by taking the advantage of both properties, we estimate the action by recognizing the trajectory, *i.e.*, evaluate whether the incoming clip is close enough to the training set.

Suppose a trajectory spanned by an action video clip be *f*, the distance from data point *x* to a curve *f* can be defined as,

$$d(x, f) = \inf_{t} \| f(t) - x \|^2 \tag{4}$$

where the curve is given as a mapping from image to R^d. The recognition of human actions can be expressed as finding the minimum average Euclidean distance to the curves of certain action data set. With a video clip has *m*-frames, and action set *j* be represented by curve f_j, the recognized action label *j** would be,

$$j^* = \arg\min_{j} \frac{1}{m} \sum_{k=1}^{m} d(x_k, f^j) \tag{5}$$

In this method, there are two parameters to be detected, the recognized trajectory *j* and the optimal offset *k*. Actually, *j** may not be the correct answer although it is best matched. To avoid the mistake, we use more reliable decision scheme.

If the trajectories represented as piece-wise linear polygonal lines connected by *n* vertices $\{v_i | i=1..n\}$, then the distance to the curve can be expressed as aligned projection distance (APD),

$$\begin{aligned} d_n(x, f) &= \min_{1 \le i \le n-1} d(x, S_i) \\ s.t.\ S_i(t) &= tv_{i-1} + (1-t)v_i,\ 0 \le t \le 1 \end{aligned} \tag{6}$$

which computes the minimum projection distance from *x* to the segment S_i.

For each incoming query video clip *C* with an unknown action, we calculate the distance between the curve of *C* and each of those clips in the training set containing *N* training samples. A distance array D_N can be generated by different action labels.

The decision scheme is quite straightforward. After the distance matrix is generated for each learning video clip, we group the distance in each action class and calculate the mean distance for each action labels. The incoming video clip is then labeled as the one which has minimum mean distance. This will be later called the mean decision method.

4. SIMULATION RESULTS

(1) Data Set

The proposed solution is tested on the KTH human activity data set which is also used in [3]. It contains 6 human actions, '*boxing*', '*handclapping*', '*handwaving*', '*jogging*', '*running*', and '*walking*'. Actions are performed by a total of 25 individuals in four different settings: S_1--out door; S_2--out door, with camera zooming; S_3-- out door, with different clothes on; S_4-- in door.

Figure 2: Human Action Clip Examples

In each setting, for every single person there are six actions and each with four video clips, with each segment's start and end frame number listed as a ground truth file. People act with different clothes, background and move to different directions in the clips. Each setting has $4 \times 25 \times 6 = 600$ actions, and the data set comprises of a total of 2400 clips. Some examples from the action clips are plotted in Figure 2. From left to right, the top row actions are, walking, jogging, and running, and the bottom row actions are boxing, hand waving and hand clapping, respectively.

The video clips are of 160×120 pixel resolution, and in processing stage, we down sample the sequence into 20×15 icon image sequences for geometry modeling.

(2) Result for Curve-Distance Approach

For recognition simulation, the classic "leave-one-out" scenario is used to test the performance. For each settings composed by 600 video clips, 24 are selected as testing data and other 576 are used for learning. Then for each test video clip, a distance matrix is generated for decision. Simulation result of the mean distance method is shown in Figure 3. The recognition result is shown in a 6×6 confusion matrix. Gray histogram in color is also used to present the value of each number: white stands for 1 while black means 0.

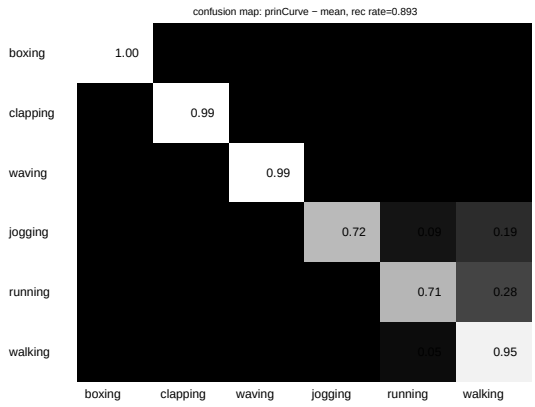

Figure 3: Human Action Recognition Performance by Principle Curve Methods

The number of diagonal line in the matrix means a correct recognition and in other places, a mistake is detected in the result. The recognition accuracy is about 89% which outperforms many existing methods in the literature.

From this results we can see that the action 1, 2 and 3 are separated nearly perfect, and the decision mistake is mainly in the class 4, 5 and 6, which are "jogging", "running" and "walking" respectively. This means the proposed methods can recognize the spatial characters but are not strong enough to keep the temporal properties.

To improve this, we combined our method with the DLFT [11] approach, which has a better performance in dealing with the temporal properties of action clips. Simulation result shows that the curve-distance approach achieves a better performance in spatial domain while the differential trajectory one is more powerful in handling temporal features.

As a result, we proposed a new combined solution: calculate the distance matrix for clip and decide whether the action should be a "hand action" (action 1, 2 and 3) or a "body action" (action 4, 5 and 6). For a hand one, principle curve method is applied while for the body action we use the DLFT. The combination result is shown in Figure 4, which achieved a recognition accuracy of 92%, better than other previous techniques.

Overall, the proposed solution achieves better performance than the state-of-the-art methods.

5. CONCLUSION AND FURTHER WORKS

In this work we proposed a curve-distance approach, which is based on the geometry information for video clips, to solve a human action recognition problem. Our solution is tested on the KTH data set and demonstrated to be effective and efficient. The performance is comparable or even better than some existing techniques in the literature. We also combined this method with other approaches to achieve a better performance.

This method is a little computational expensive for a real-time application because of the huge amount of calculation in distance matrix generation. But, it is solvable. Our future work will be mainly focused on simplifying the process. Fast algorithms will be of great significance.

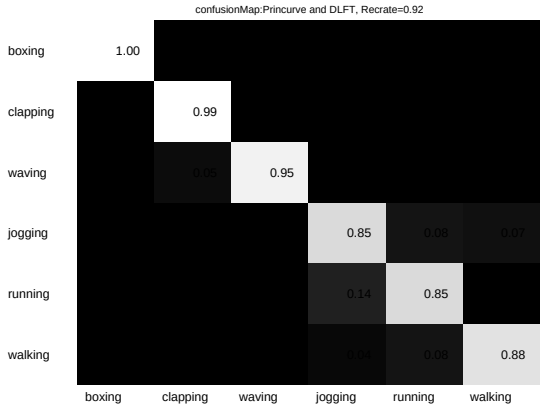

Figure 4: Recognition Performance Confusion Matrix: Combination of Principle Curve and DLFT methods

REFERENCES

[1] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A Fully Automated Content-Based Video Search Engine Supporting Spatio-Temporal Queries" , *IEEE Trans on Circuits and Systems for Video Tech* (CSVT), vol. 8, no. 5, pp. 602-615, September 1998.

[2] L. Laptev and T. Kindeberg, "Space-time interest points", *Proc. of IEEE ICCV*, pp. 432-439, Nice, France, 2003.

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach", *Proc. of IEEE ICPR*, pp. 32-36, UK, 2004.

[4] Y. Ke, R. Sukthankar, and M. Hebert. "Efficient visual event detection using volumetric features", In Proc of IEEE ICCV, pp. 166- 173, 2005.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei. "Unsupervised learning of human action categories using spatial-temporal words", *Proc of British Machine V*ision Conference, Edingburg, 2006.

[6] Z. Li, L. Gao, A. K. Katsaggelos, "Locally Embedded Linear Subspaces for Efficient Video Indexing and Retrieval", *Proc. of IEEE Int'l Conf on Multimedia & Expo*, pp. 1765-1768, 2006.

[7] T.-Y Kim, S.-F. Wong and R. Cipolla, "Tensor Canonical Correlation Analysis for Action Classification", *Proc. of IEEE Conf. on CVPR*, Minneapolis, MN, 2007

[8] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60, 2 (2004), pp. 91-110.

[9] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linearprojection", *IEEE Trans on PAMI*, vol. 19, no. 7, pp. 711-720, July 1997.

[10]L.Smith, "A tutorial on Principle Component Analysis", [online] http://kybele.psych.cornell.edu/%7Eedelman/Psych-465-Spring-2003/PCA-tutorial.pdf.

[11] Z. Li, Y. Fu, S. Yan, and T.S. Huang, "Real-Time Human Action Recognition by Luminance Field Trajectory Analysis", *ACM Multimedia*, 2008.