Real-Time Human Action Recognition by Luminance Field Trajectory Analysis

Zhu Li
Dept of Computing
Hong Kong Polytechnic University
Kowloon, Hong Kong
+852 2766-7316
zhu.li@ieee.org

Yun Fu, Thomas S. Huang
Dept of ECE
University of Illinois at Urbana-Champaign, USA
+1 217-244-2960
{yunfu2,huang}@ifp.uiuc.edu

Shuicheng Yan
Dept of ECE
National University of Singapore
Singapore
+65 6516-2116
elesyan@ece.nus.edu.sg

ABSTRACT

The explosive growth of video content in recent years fueled by the technological leaps in computing and communication has created new challenges for video content analysis that can serve applications in video surveillance, video searching and mining. Human action detection and recognition is one of the important tasks in this effort. In this paper, we present a luminance field manifold trajectory analysis based solution for human activity recognition, without explicit object level information extraction and understanding. This approach is computationally efficient and can operate in real time. The recognition performance is also comparable with the state of art in comparable set ups.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Video

General Terms

Algorithms, Performance, Experimentation.

Keywords

Video Analysis, Video Understanding, Human Action Recognition.

1. INTRODUCTION

.

The advances in computing and communication technologies make the capture, communication and storage of video content easier and cheaper than ever. To enable more intelligent and meaningful applications with this vast amount of video data, especially in surveillance and on-line video repository searching and mining, an efficient and effective video content analysis and learning solution that can detect and recognize human actions is becoming a crucial task in the effort of making video content more accessible and intelligent. Example applications include surveillance video analysis for security, sports video analysis for labeling and searching.

Vast amount of research work exists in this area. Typical approaches include object segmentation and motion estimation and based solutions, [Chang98][Zelnik01][Yilmaz05], which try to explicitly recover object and motion information and then recognize human action based on learning of object level spatio-temporal primitives. This approach suffers from poor robustness to the appearance variances of human actions with different subjects , lighting, background conditions and occlusions.

To address this issue, spatio-temporal interests points [Laptev03] are detected to avoid the robustness issues in object level detection, segmentation and tracking. The human actions therefore are recognized from spatio-temporal points primitives via local SVM in [Schuldt04], and generative graphical models in [Niebles06].

Instead of interests points, which are pre-selected localized features that are suppose to be robust to image formation variations, the spatio-temporal volume modeling based solutions are also investigated in [Ke05] and [Kim07]. In [Kim07] the feature selection is integrated into the Tensor CCA learning process.

In a more general video clip search problem, a query clip is given and a system is to find the repeats of the same clip some image formation variations and signal degradations. Again, signal and object level features and statistics can be utilized to create a matching criterion. In our previous work [Li06] [Fu08], a luminance field trajectory modeling and matching solution is proposed and the resulting video clip searching achieves very high performance in both precision-recall and response time. The key observation there is that the luminance field trajectory of video sequences contains sufficient information for a variety of detection and recognition problem, and can be implemented efficiently for real-time operations.

In this paper we extend our luminance field analysis work to tackle the task of real-time human action recognition by video analysis. The video input is represented by its luminance field trajectory, and its geometric features contains discriminating information for human action recognition, which can be extracted by an efficient algorithm.

The paper is organized into the following sections. In section 2, we present the luminance field trajectory extraction and modeling solution. In section 3 we discuss our recognition algorithm based on trajectory geometry feature learning, and in section 3, we present simulation results and discuss the performance of the proposed algorithm, and finally we draw conclusion and outline future work in section 4.

2. Luminance Field Trajectory

Video clips can be viewed as some trajectories in a high dimensional space. Considering a clip of *n* frames with luminance field size of *W* x *H* pixels, each frame F_k can be represented as a point in some space R^{WxH}.

$$F_k \in R^{WxH}, \forall k \in [1..n] \quad (1)$$

Video clips containing different human activities will have different trajectories in this space and the geometry of these trajectories contains sufficient statistics to recognize different human actions. A computationally practical trajectory analysis and discriminative modeling is the key to the task.

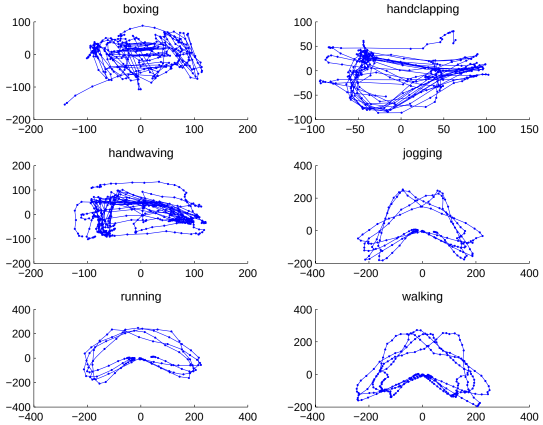


Figure 1. Human Action Luminance Field Trajectory Examples

Obviously, the original $R^{W \times H}$ space contains more information than necessary and helpful to our tasks. The dimensionality is too high for effective modeling. Let the trajectory be obtained through,

$$x_k = AF_k = \left[a_1^T, a_2^T, \ldots, a_{WxH}^T\right] F_k \quad (2)$$

where the subspace *A*: *d* x (*W*x*H*) is obtained through an unsupervised local learning, with the objective of preserving the maximum amount of information while has a reasonable number of dimensions, *d*, for subsequent modeling and learning algorithm to be effective. Naturally we chose localized PCA to find the subspace *A*,

$$\begin{aligned} a_1^T &= \underset{a}{arg\,max}\, \underset{F_k}{var}\left(a^T F_k\right) \\ a_2^T &= \underset{a}{arg\,max}\, var\left(a^T \tilde{F}_1\right) \\ &\ldots\ldots. \\ a_k^T &= \underset{a}{arg\,max}\, var\left(a^T \tilde{F}_{k-1}\right), \end{aligned} \quad (3)$$

The iteration on subspace bases are based on the data,

$$\tilde{F}_{k-1} = F_k - \sum_{j=1}^{k-1} a_j a_j^T F_k \quad (4)$$

Some example 2D luminance field trajectories of human actions from the KTH data set [Schuldt04] are plotted in Fig. 1. All 6 human actions in consideration, boxing, hand clapping and hand waving are plotted.

An intuitive solution to recognize human actions from these trajectories is to cast it as a shape or trajectory matching problem. Unfortunately, the translation, scale and rotation invariance issues are not easy to handle in this case. Instead, we model the differential trajectories of different human actions and try to apply learning in differential trajectory space.

3. Differential Trajectory Based Learning

Let the differential trajectory of video clip be,

$$d_k = \begin{cases} 0, & k = 1 \\ \left\|x_k - x_{k-1}\right\|, & k > 1 \end{cases} \quad (5)$$

Now we have an 1-D trace representation of different human action clips, but obviously we lose information in the process. We will demonstrate later that this differential trace still contains enough information for action recognition, while can be computed efficiently in real-time.

Some examples of differential traces are plotted in Fig. 2,

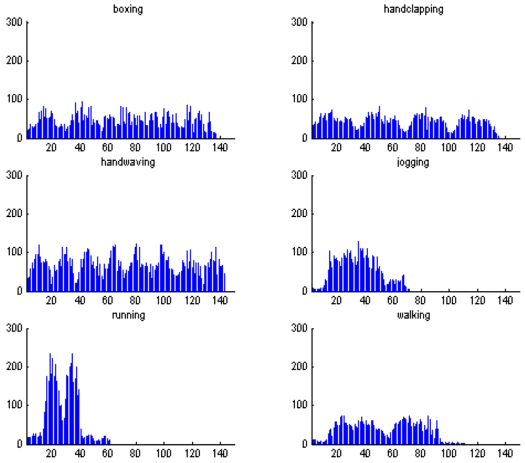


Figure 2. Differential Trace Examples

Notice that in real-time applications, the duration of action clips can be also different. Therefore direct matching in differential trace domain is not desirable. Instead, the differential trajectory is represented in the frequency domain via the Discrete Cosine Transform (DCT). .

An 64-point DCT is performed on a sliding time window over the differential trace. This will incur a roughly 2-sec delay for real time, or on-line applications, which is acceptable in most cases.

Some example sliding window differential trace DCT features are plotted in Fig. 3. Notice that only the first d_{DCT}=32 DCT coefficients are plotted, since most high frequency coefficients thereafter are zeros. The plots are obtained by averaging the DCT coefficients over different subjects and camera settings. Notice that certain patterns can already be understood in this plot.

The process so far can reduce each segment of 64 frames in video sequence to a differential trajectory in some local PCA space, and then represent it as a d_{DCT} dimensional vector. The action labels are known from the training set, therefore we can apply supervised learning here for the recognition.

Standard supervised learning techniques like LDA in Fisherface [Belhumeur97], as well as more recent graph embedding based Locality Preserving Projection (LPP) [He05] are applied to these feature vectors in the DCT domain.

The choice of DCT length should also offer enough time resolution that matches the smallest time scale to recognize certain action. In this work, as target actions are *boxing*, *hand clapping*, *hand waving*, *running*, *jogging* and *walking*, 2 seconds is roughly the bare minimum to allow for the system to pick up enough features for correct recognition.

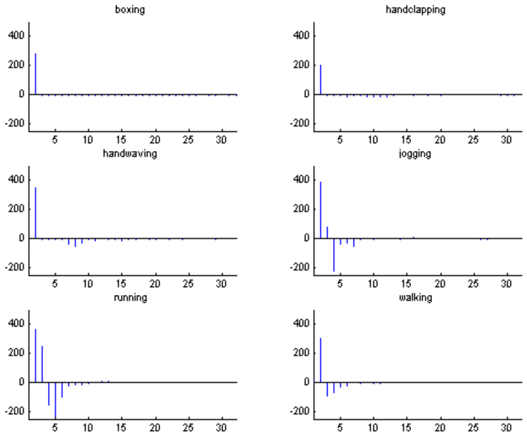


Figure 3. DCT features of differential traces

Computationally, the on-line feature extraction part only involves buffering 64 frames in a sliding window, compute local PCA space trajectories via scaling and projection, differentiation and then DCT. This can be done in real time with a minimum burden on typical computing devices nowadays.

The classification is based on a simple Gaussian Mixture Model (GMM) classifier. The training clips projected in LPP or LDA space are modeled as a 6-class Gaussian mixture, and the classification is done by assigning action labels that gives maximum likelihood.

We have also experimented with more sophisticated classifier like Support Vector Machine (SVM), but the benefit in performance is limited at this time. Simulation results and discussions are presented in the following section.

4. SIMULATION RESULTS

4.1 Data Set

To test the algorithm developed, we use the KTH human action data set [Schuldt04], which contains 6 human actions, '*boxing*', '*handclapping*', '*handwaving*', '*jogging*', '*running*', and '*walking*'.

Actions are performed by a total of 25 subjects in 4 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.

For each setting, each action has 4 video clips, with each segment's start and end frame number listed as a ground truth file. Each setting will have 4 x 25 x 6 = 600 action clips of varying durations, and the data set comprises of a total of 2391 clips, with a small number of entries missing.

The video clips are of 160 x 120 pixel resolution, and in pre-processing stage, we down convert the sequence into 20 x 15 icon image sequences for trajectory computation. The localized PCA is therefore trained with sample clips in R^{20x15}, with d=12 used in trajectory representation.

Some examples from the action clips in [Schuldt04] are plotted in Fig. 4. 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.

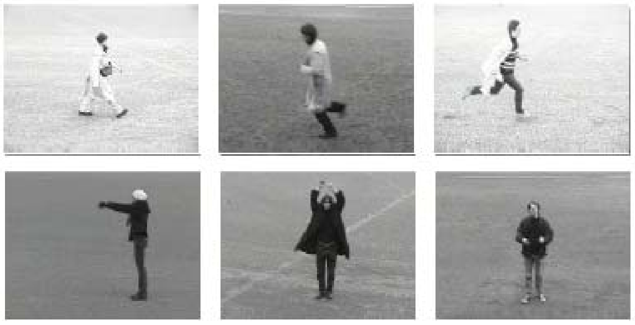


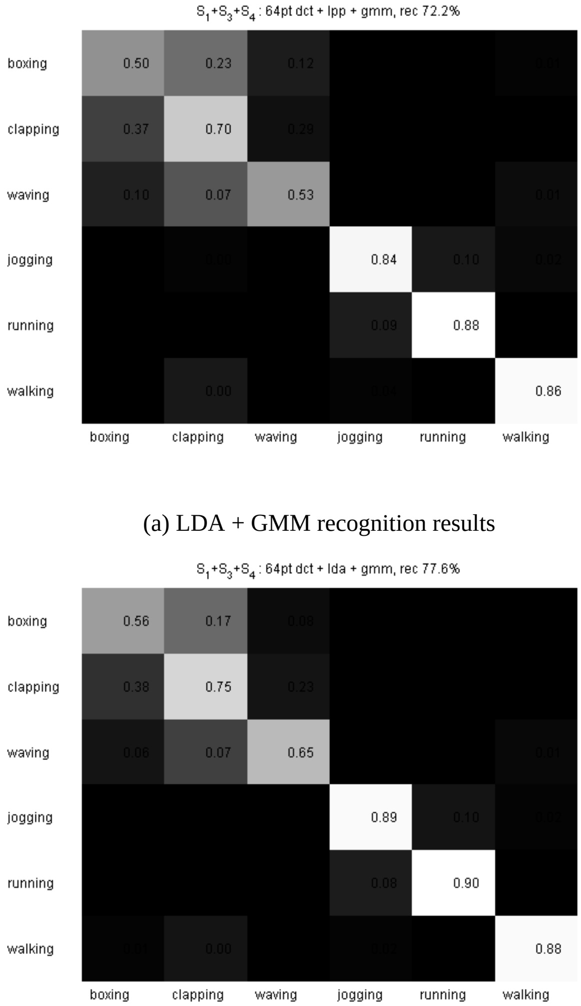
Figure 4. Human Action Clip Examples

Notice that there are a variety of lighting conditions and the subject also wearing different clothes.

4.2 Discriminant Features Learning

After obtaining the differential trajectories for each video clip, the supervised learning is applied. , and DCT over a sliding window of size $W=64$, we obtain a feature vector of d_{DCT} = 12 for the subsequent supervised learning process. The cut of at first 12 coefficients are selected from simulation, since DCT coefficients after AC11 are mostly zero and do not contribute much to the learning.

For each iteration of the recognition experiment, the data set is partitioned into non-overlapping training set of 500 clips and the test set of 100 query clips. The recognition accuracy performance results are then obtained by averaging over 10 iterations. The confusion maps are plotted in Fig. 5a for the LPP+GMM case, and Fig. 5B for the LDA+GMM case.



(a) LDA + GMM recognition results

(b) LPP + GMM recognition results

Figure 5. Human Action Recognition Performance

The 64-point DCT features with LPP learning achieves an average of 72.2% accuracy in recognition, for settings S_1, S_3, and S_4, while the LDA based learning achieves a better accuracy of 77.6%. The performance is comparable or better with the performances in [Schuldt04], and [Ke05], but not in [Kim07], mainly because an important pre-processing step of box cropping of human figure out of video sequence is not applied.

5. Conclusion and Future Work

In this paper we presented a real-time human action recognition solution that is based on luminance field trajectory analysis and learning. The simulation results demonstrated the effectiveness of the proposed solution in recognition accuracy. The solution is also computationally simple, and can operate efficiently with a sliding window over the observed video sequences in real time without off-line processing with long delays.

In the future we will investigate more sophisticated pre-processing steps like human figure cropping, as well as better classifiers than simple GMM. Parameters will also be further fine tuned to make the proposed solution more efficient and effective.

6. REFERENCES

[Chang98] 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.

[Fu08] Y. Fu, Z. Li, T. S. Huang, and A. K. Katsaggelos. “Locally Adaptive Subspace and Similarity Metric Learning for Visual Clustering and Retrieval”, *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 390-402, 2008.

[Ke05] Y. Ke, R. Sukthankar, and M. Hebert. “Efficient visual event detection using volumetric features”, In *Proc of IEEE ICCV*, pp. 166- 173, 2005.

[Kim07] T.-Y Kim, S.-F. Wong and R. Cipolla, “Tensor Canonical Correlation Analysis for Action Classification”, In *Proc. of IEEE Conf. on CVPR*, Minneapolis, MN, 2007

[Laptev03] L. Laptev and T. Kindeberg, “Space-time interest points”, *Proc. of IEEE ICCV*, pp. 432-439, Nice, France, 2003.

[Belhumeur97] 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.

[Li06] 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.

[He05] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, “Face Recognition Using Laplacianfaces”, *IEEE Trans. on PAMI*, vol. 27, no. 3, pp. 328- 340, Mar. 2005.

[Niebles06] J. C. Niebles, H. Wang, and L. Fei-Fei. “Unsupervised learning of human action categories using spatial-temporal words”, *Proc of British Machine Vision Conference*, Edingburg, 2006.

[Schuldt04] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing Human Actions: A Local SVM Approach”, *Proc. of IEEE ICPR*, pp. 32- 36, Cambridge, UK, 2004.

[Yilmaz05] A. Yilmaz and M. Shah. “Recognizing human actions in videos acquired by uncalibrated moving cameras”, *Proc. of IEEE ICCV*, vol. 1, pp.150-157, 2005.

[Zelnik01]L. Zelnik-Manor and M. Irani, “Event-based analysis of video”, *Proc. of IEEE CVPR*, 2001.