Unsupervised approaches to visual analysis of human motion: towards automatic classification of activity and behavior

Roberto Pugliese, Jayasimha Rao, and
Santosh Tirunagari

Aalto University School of Science
Department of Information and Computer Science
roberto.pugliese@aalto.fi,
{tsantosh7,jayasimha.ramachandrarao}@gmail.com
P.O. Box 15400, FI-00076 Aalto, Finland
http://www.ics.tkk.fi

Abstract. Human motion analysis is one of the most active research areas in computer vision due to the undercurrent and potential applications such as visual surveillance, perceptual user interface, content-based image storage and retrieval, video conferencing, athletic performance analysis, virtual reality, etc. It is clear how crucial it is for a human motion analysis system to understand different activities performed by the human. The problem is by no means trivial since it involves many steps such as human detection, pattern recognition and understanding of context. In this paper we focus on the problem of classifying different activities and behaviors of a human and concentrate on machine learning techniques that do not require any prior knowledge (unsupervised) of human motion. To further narrow down the scope of this paper we will assume the system is able to track human motion and start our survey from there on. Motion tracking is a mature field and commercial systems are available. We review an implementation of Unsupervised classification of motion sequences into different actions using Aligned Cluster Analysis (ACA) and other kernel methods. The system is able to classify different human motion (walking, running, jumping, etc.) from a long motion tracking file. We evaluate quantitatively the effectiveness of the implementation for the chosen dataset, addressing the limitations encountered. From there we draw some conclusions and recommendation for concrete applications of machine learning techniques in the field of bodily human-computer interaction.

Keywords: clustering, ACA, k-means, unsupervised learning, human, behavior, motion

1 Introduction

The field of ubiquitous computing proposes a future where technology will gradually be embedded in the daily life objects, in the environments we live in as well

as in portable smart devices. Within this prediction, the role of technology in understanding and accompanying human activities is essential in many areas of application. The approach of HCI has shifted moving away from computer-centered designs toward human-centered designs for HCI, made for humans based on models of human behavior [6]. According to these views, human-centered designs should focus on understanding what is communicated (linguistic message, nonlinguistic conversational signal, emotion, attitude), how the information is passed on (the persons facial expression, head movement, nonlinguistic vocalization, hand and body gesture), why, that is, in which context the information is passed on (where the user is, what his or her current task is, are other people involved), and which (re)action should be taken to meaningfully respond to the user. The goal has wide range of potential applications such as smart surveillance, motion based diagnosis, and advanced user interface. The latter includes nowadays home-entertainment and gaming based on different sensor technologies introduced by the gaming console manufacturers.

1.1 Context: Behavior Analysis

Capturing the complexity of human-behavior is a hard challenge far from being solved. In this paper we focus on a computer vision based techniques of motion and behavior analysis. Thus limiting the problem to the understanding of human behaviors, from image sequences involving humans. Three major issues are involved in this process:

- Human detection
- Tracking
- Activity understanding

The goal of the activity understanding is to analyze and interpret human action and the interactions between people and other objects.

Figure 1 show a general framework for human motion analysis propose by Wang et al in [4]. The process involves acquiring increasingly higher-level knowledge of the scene encoded as a sequence of images. For the purpose of this paper, the action recognition and emergence of semantic are discussed and ad-hoc machine learning techniques described.

1.2 Human Sensing

In the last few years the proliferation of so-called natural user interfaces such as Nintendo WiiMote, Sony PlaystationMove and Microsoft Kinect not just promised a renewed importance of bodily involvement in the gaming experience but also opened the door to designers and hobbyists experimenting with bodily interaction in different research contexts. Though only in particular settings and contexts, body tracking with and without markers is relatively simple and robust. In the case of the Microsoft Kinect sensor, the system is able to track up to four human bodies with decent accuracy for non-critical human-computer

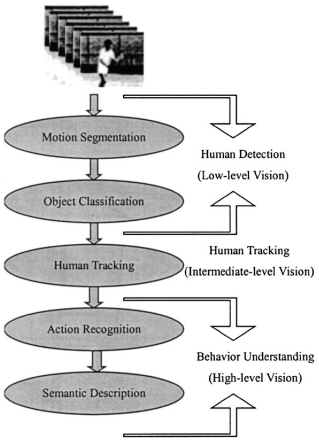

Fig. 1. A general framework for human motion analysis, from Wang, L., Hu, W. & Tan, T., 2003.

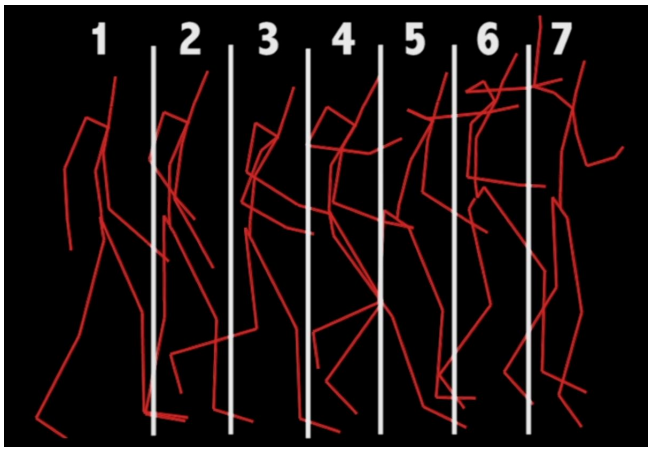

Fig. 2. KINECT motion sequence.

interaction. The position of the main joints of the human body is computed at 30 frames per second shown in the figure 2.

For this reason we will assume the problem of object classification and human body tracking solved. The reader can imagine of a system able to recognize presence of human bodies in a scene and able to represent them as stick figures alike the one shown in Figure 3.

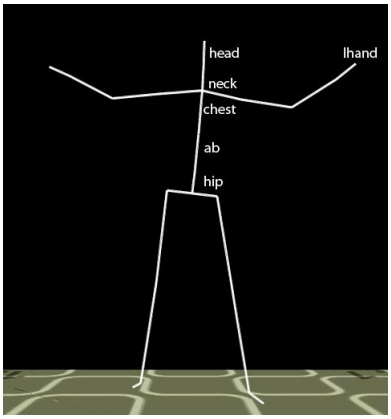

Fig. 3. The human body represented as a stick figure. Upper-body joints labels are reported.

1.3 Organization of the Paper

In the reminder of this paper, the problem of activity understanding is addressed by means of machine learning methods. In particular in Section 2, unsupervised methods are proposed for the classification of different activities from a long motion sequence. Results of an implementation from other authors and the application of the our method to the data set of a recoded motion are presented in Section 3. In Section 4 the problem of predicting behavior from learned spatio-temporal patterns is addressed. Two methods, namely hidden-markov models and vector quantization are presented. Finally, Section 5 analyzes some scientific challenges and presents some possible directions for future research.

2 Methods

In this section, we provide a brief description of the different methods that could be used for temporal segmentation of video sequences and methods that can determine the object behaviors.

2.1 Temporal segmentation

First we concentrate on unsupervised learning techniques for temporal segmentation, and describe behavior classification techniques.

Kernel K-Means

K-means clustering owing to its simplicity, is the most popular unsupervised learning technique. It splits a set of n data points into k clusters so that the inter-cluster similarity is maximized. To put it in a more mathematical way, it works by optimizing the energy function

$$J_{km} = \sum_{c=1}^{k} \sum_{i=1}^{n} g_{ci} ||d_i - m_c||_2^2 \tag{1}$$

such that $G^T 1_k = 1_n$ and $g_{ij} \in 0, 1$, where $d_i \in \mathcal{R}^{d \times 1}$ is the data vector. g_{ci} is the indicator variable, i.e $g_{ci} = 1$ is d_i is in cluster c.

K-means though simple and effective, has its limitations. It works best on spherical data. For non-spherical data, the partitions obtained are not optimal. One way of overcoming this is to map the data into a higher dimension using a kernel, and use k-means in the mapped space. Kernel K-means works by optimizing a similar equation to Eq.1 but in a higher dimension.

$$J_{kkm} = \sum_{c=1}^{k} \sum_{i=1}^{n} g_{ci} ||\phi(d_i) - \phi(m_c)||_2^2 \tag{2}$$

where $\phi(.)$ is the mapping function. The distance function $dist_c(d_i) = ||\phi(d_i) - \phi(m_c)||_2^2$ is given by

$$dist_c(d_i) = \kappa_{ii} - \frac{2}{n_c} \sum_{j=1}^{n} g_{cj} \kappa_{ij} + \frac{1}{n_c^2} \sum_{j_1, j_2 = 1}^{n} g_{cj_1} g_{cj_2} \kappa_{j_1 j_2} \tag{3}$$

where κ_{ij} is the kernel function.

Aligned cluster analysis

Aligned cluster analysis (ACA) is an extension of the Kernel K-means. It partions the given sequence $X \in \mathcal{R}^{d \times n}$ into m disjoint segments. Each of the segments correspond to one of k actions. The segment $Y_i = X_{[s_i, s_{i+1})}$ is a sequence of frames beginning at s_i and end at $s_{i+1} - 1$.

ACA obtains the partions by optimizing the energy function:

$$J_{ACA}(G, s) = \sum_{c=1}^{k} \sum_{i=1}^{m} g_{ci} dist_c(X_{[s_i, s_{i+1})}) \tag{4}$$

Kernel K-means and ACA differ from each other in two respects: One, ACA can cluster sequences of different lengths, while in Kernel K-means, the sequence

lengths are fixed. Secondly, ACA uses Distance time warps (DTW) to calculate the kernel $dist_c(Y_i)$. DTW is known to be robust to noise and invaiant to the speed of the action.

However, DTW is not a properly defined metric, as it does not satisfy the triangular inequality property. *Dynamic Time Alignment Kernel*(DTAK) proposed by Shimodaira et al. [7], which is a metric between two time sequences can be used instead of a DTW kernel. It is defined between two time sequences $\hat{X} = \{\hat{x_1}, \hat{x_2}, ..., \hat{x_{n_1}}\}$ and $\tilde{X} = \{\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_{n_2}}\}$ as:

$$\kappa(\hat{X}, \tilde{X}) = \frac{p_{n1,n2}}{n_1 + n_2}$$

$$p_{i,j} = max \begin{cases} p_{i-1,j} + \kappa_{ij} \\ p_{i-1,j-1} + 2\kappa_{ij} \\ p_{i,j-1} + \kappa_{ij} \end{cases}$$

where $\kappa_{ij} = e^{\frac{-1}{2\sigma^2}||\hat{x_i} - \tilde{x_j}||^2}$ and $p_{0,0} = 0$.

2.2 Behavior Classification

Once the given video sequence is segmented into different actions, we can use this information to determine object behavior. Based on contextual information, we try to build a model for the object behavior which would not only account for observed behavior, but also be able to predict future actions. In the rest of this section, we describe two methods which try to accomplish this. First a Hidden Markov Model(HMM) based method, and then a method which uses a Neural Network approach to solve the problem.

Hidden Markov Model

Hidden Markov model (HMM) based techniques have been successfully used for predicting patterns, specially in speech recognition research. Similar to an n-gram model, we can construct a HMM based model which based on the previous n actions, would be able to predict the $n + 1$th action. [8] have sucessfully used HMM to predict which room a person would enter next based on previous n rooms he has visited. The same technique can be applied to the problem at hand, in that we try and predict what would be the next action given previous n actions.

Vector Quantization

[9] have proposed a neural network based approach to modelling behavior using context information. It consists of two competitive learning neural networks which are connected using a leaky integrator. Figure 4 shows the network architecture used for modeling object behavior.

The Symbol Network models the the complex probability density function for the input data $x \in \mathcal{R}^n$ at any given time. This is done using Vector Quantization

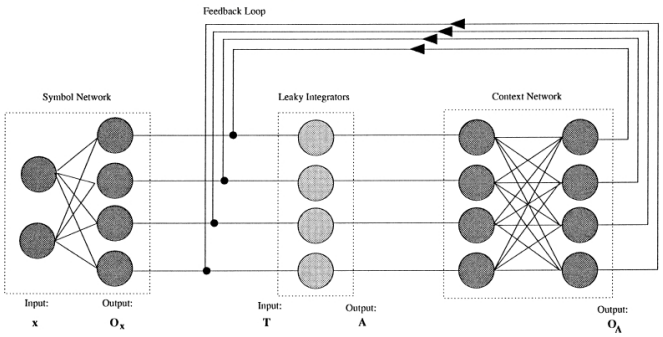

Fig. 4. Approach used by [9] as a network architecture.

(VQ). VQ works by obtaining a representation of the feature space using a number of prototype vectors. These prototype vectors are obtained using competitive learning.

The context Network models the probability density function of object behaviors. Again VQ is used to learn the appropriate weights for each of the synapses. The two networks are connected to each other using a leaky integrator, which is a neuron which stores the information of its previous activation. Finally it has a feedback loop, which helps the network to learn the relationship between the activation trace and the next input vector.

3 Experiments

In this section, we describe the experiments on motion capture data and KINECT data. We compare the segmentation performance of ACA with kernel k-means algorithm.

3.1 Motion capture Data

Motion capture data has 15 sequences performed by subject 86, each of which is a combination of 10 natural actions (e.g. walking, punching, drinking, running). Typically each sequence contains 8000 frames (70 secs).

3.2 Implementation and Results

We used unsupervised learning methods. Hence we performed the experiments using Kmeans algorithm for clusters ranging from 2 to 14. The accuracy graph is taken to analyse which cluster gave good performance. We had 10 natural actions and the accuracy curve gave us the same. We analysed the temporal segmentation of the motion capture data using chisquare and linear kernel kmeans algorithm and compared it with the ACA and HACA and GMM algorithms.

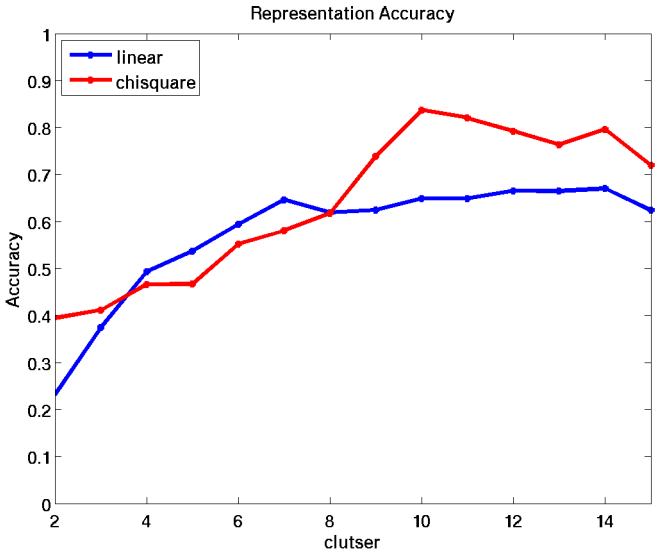

Fig. 5. Cluster representation accuracy.

The convergence of the linear kernel kmeans was very slow when compared to chisquare kernel.

from the figure 5, 10 clusters gave performace and we chose 10 clusters for temporal segmentation of the motion capture data. Chisquare kernel gave good performace compared to linear kernel. The best accuracy obtained was 0.8 with the chisquare kernel kmeans. We have also conducted the experiments with ACA, HACA and GMM algorithms aswell.

From the figure 6, the top graph shows the ground truth. ACA and HACA gave the good results with 0.9 and 0.91 accuracy, when compared to GMM wih 0.8 accuracy. From these results ACA and HACA are the best algorithms to capture the temporal segmentation of the human motion.

4 Conclusions

The problem of understanding and intelligently accompany human activity is at the core of human-computer interaction. It necessitates human-centered designs to focus on understanding human intention, context and meaningful way to reacting or responding to human behavior. In this paper, we focused on how computer vision-based techniques of motion tracking can be complemented by different machine learning method in order to automatically classify and predict human activity. An important topic for future work is to explore how a machine could learn new schemas (i.e., concept learning), perhaps through a process of discovering new clusters of interaction patterns. The unsupervised

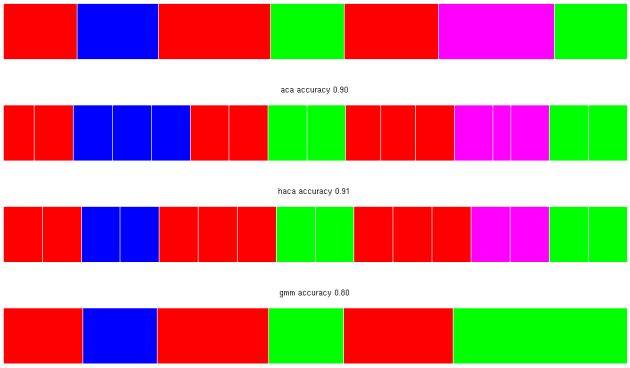

Fig. 6. Accuracies comparison between ACA, HACA, GMM to the groundtruth.

approach showed seems suitable for an on-line implementation that will increase the number of concepts learn by the machine in the history of its observation. Moreover, it is worth observing human behavior embraces high-level semantic events, which typically include interactions with the environment and causal relationships. Present approaches to machine analysis of human behavior are neither multimodal, nor context-sensitive. For that, the future of HCI is tied to research about multi-modalities of the sensing system. Humans simultaneously employ modalities of sight and sound and an intelligent system needs to be equipped with multiple senses for multisensory concept formation. Challenges arise from how to model the fusion of the different modalities, and how to operate context-dependent fusion and discordance handling. Finally, not just the context but also the temporal evolvement of the behavioral cues play a crucial role in machine analysis of human behavior.

References

1. An efficient k-means clustering algorithm: Analysis and implementation, T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (2002), 881–892.
2. Goncalves, L. & Perona, P., 2003. Unsupervised learning of human motion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(7), p.814-827.
3. Torre, F.D., Hodgins, J.K. & Zhou, F., 2008. Aligned Cluster Analysis for Temporal Segmentation of Human Motion Aligned Cluster Analysis for Temporal Segmentation of Human Motion. 8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008. FG 08., p.1-7.
4. Wang, L., Hu, W. & Tan, T., 2003. Recent developments in human motion analysis. In Pattern Recognition. pp. 585-601.
5. Wang, T.S. et al., 2001. Unsupervised analysis of human gestures. Artificial Intelligence, p.174181.

6. Pantic, M. et al., 2006. Human computing and machine understanding of human behavior: a survey. In Proceedings of the 8th international conference on Multimodal interfaces. ACM, pp. 239248. Available at: http://portal.acm.org/citation.cfm?id=1181044 [Accessed December 12, 2011].
7. H. Shimodaira, K.I. Noma, M. Nakai, & S. Sagayama. Dynamic time-alignment kernel in support vector machine. In NIPS, pages 921928, 2001.
8. Gellert, A. (2006). Person Movement Prediction Using Hidden Markov Models. Elements, 15(1), 17-30.
9. Sumpter, N., & Bulpitt, A. (2000). Learning spatio-temporal patterns for predicting object behaviour. Image and Vision Computing, 18, 697-704.