Unsupervised Analysis of Human Gestures

Tian-Shu Wang[1], Heung-Yeung Shum[2], Ying-Qing Xu[1], and Nan-Ning Zheng[1]

[1] Artificial Intelligence and Robotics Lab,
Xi'an Jiaotong University, Xi'an 710049, P.R.China
{tswang,nnzheng}@aiar.xjtu.edu.cn
[2] Microsoft Research, China,
No49, Zhichun Road, Haidian District, Beijing 100086, P.R.China
{yqxu,hshum}@microsoft.com

Abstract. Recognition of human gestures is important for analysis and indexing of video. To recognize human gestures on video, generally a large number of training examples for each individual gesture must be collected. This is a labor-intensive and error-prone process and is only feasible for a limited set of gestures. In this paper, we present an approach for automatically segmenting sequences of natural activities into atomic sections and clustering them. Our work is inspired by natural language processing where words are extracted from long sentences. We extract primitive gestures from sequences of human motion. Our approach contains two steps. First, the sequences of human motion are segmented into atomic components and clustered using a Hidden Markov Model. Thus we can represent the original sequences by discrete symbols. Then we extract lexicon from these discrete sequences by using an algorithm named COMPRESSIVE. Experimental results on music conducting gestures demonstrate the effectiveness of our approach

1 Introduction

Recognition of human gestures is important for human-computer interfaces, automated visual surveillance, and video library indexing[1]. This process, however, involves significant problems. Typically a large collection of training examples of gestures must be acquired in order to build models for the gestures. To obtain the training examples, a substantial number of gesture sequences must be segmented and aligned, typically by hand [2]. The common practice of manual segmentation and labeling is labor-intensive and error-prone. Worse, for many challenging applications, the set of gestures is not known in advance. In this paper, we present an approach for automatically segmenting and labeling a continuous sequence of human gestures.

Our approach makes no assumption about the presence of facilitative side information such as obvious segment points or the duration time of each gesture. Instead, we consider a sequence of human activities to consist of repetitive gesture primitives with a high-level structure controlling the temporal ordering. This is analogous to the concept of words and grammar in natural language processing. The activities are exhibited by, for example, dance, Tai-chi, and sign language.

Human gestures are expressive human body motions, which generally contain spatial and temporal variation. To handle the variation, we need choose an appropriate representation. In one previous work, gestures are regarded as trajectory curves in a configuration space [3]. In our approach, we emphasize the dynamical part of gestures. We choose Hidden Markov Models (HMMs) to represent the dynamics. It has been demonstrated in [2] [4] that HMMs are effective for human gesture recognition.

The approach used in this paper is described as follows. The first step is to form a discrete representation of gestures. The observed continuous sequence is automatically segmented into atomic gestures. It is an over-segmentation process. And we use HMM models to separate atomics into several clusters. Thus by using cluster labels to replace atomics in the original sequence, we transform the continuous observation to a discrete symbol sequence. In the second step, we learn "words" from the discrete representation. Borrowing methods from natural language processing and data compression, we obtain structure from the symbol sequence and thus determine appropriate primitive gestures and labels of the original sequence. Although the computer is unaware of the meanings of the primitive gestures, the original sequences are effectively represented by them. For example, if we correspond word 16 to a waving hand gesture, we know that all other positions in the original sequence labeled by word 16 are waving hand.

In the reminder of this paper, we briefly overview related works in section 2, describe the details of our approach in section 3, present experiment results in section 4, and end with discussion and future work in section 5.

2 Related works

A vast amount of work in gesture recognition has been performed in the area of computer vision, and is reviewed in [4]. These works can be divided into two categories: trajectory-based and dynamics model-based. The trajectory-based approach matches curves in configuration space to recognize gestures [3]. The dynamics model-based approach learns a parametric model of gestures. HMM is a typical dynamics model and was proven to be robust in its recognition of gestures [2]. The HMM model has been extended to a more general model named Dynamic Bayesian Networks [5].

Several works involve unsupervised learning of video sequences and gestures. A HMM-based approach is used to cluster ambulatory audio and video content [6]. In it, the number of clusters must known a priori. An entropy training process of HMM is proposed in [7] and used to learn office activity. The whole sequence is used to train a single model, thus it is hardly suitable for handling large scale problems. An incremental learning framework of natural gestures is proposed in [8], but it does not involve learning the high level structure of gestures.

A relevant topic to this paper is learning the lexicon of a natural language. A classical MK10 algorithm is used to infer word boundaries from artificially

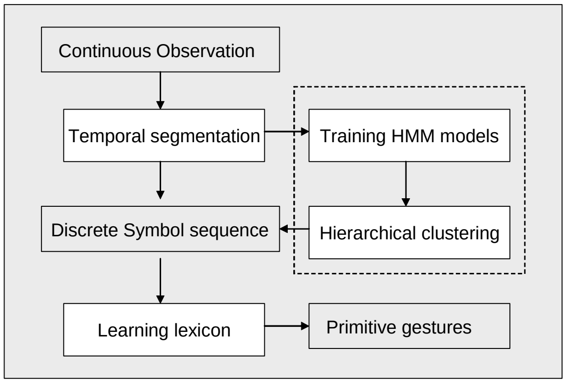

Fig. 1. Framework of the unsupervised learning approach

generated natural language sentences in [10]. The sequitur algorithm is used to build hierarchical structure in an online process with linear calculation complexity in [9]. A dynamical programming-based approach for extraction of lexicon is proposed in [11]. The performance of different algorithms are not easily compared, but in terms of learning words, an offline algorithm is better than an online algorithm.

3 Approach

Human gestures can be represented as a sequence of hand positions in 3d-space. In a video sequence, the 3d space-curve is projected to a 2d image plane as a 2d trajectory. Thus the continuous observation of gestures is an ordered sequence of hand positions in a 2d image plane.

Figure 1 exhibits a framework of the learning approach. First, the continuous observation of human gestures is segmented into atomics movements. The segmentation involves identifying suitable break points at which to partition the gestures. The result of this process is an over-segmentation of gestures, in which every segment is only an atomic movement and without much meaning. Then we cluster those segments into several clusters by using Hidden Markov Models. This is done to learn a HMM for each segment using a hierarchical clustering method. The result of clustering gives a discrete representation of the original continuous observation, in which every segment is replaced by the cluster number that it belongs to. Finally, we infer the lexicon from the discrete symbol sequence. The details of each step are described in the following.

3.1 Temporal segmentation

The purpose of temporal segmentation is to split the continuous sequence into atomic segments. The atomic segments exhibit basic movements whose execution is consistent and easily characterized by a simple trajectory. As we need to extract meaningful gestures, the segmentation is overly fine and can be considered as finding the alphabet of motion.

The segmentation involves searching natural inconsistent points within the whole observation. A change in the type of human movement usually causes dips in velocity or abrupt variations in moving direction. We exploit this by finding the local minima of velocity and local maxima of change in direction. The minima (maxima) below (above) the certain threshold are selected as segment points. In practice, we found the calculation of change in direction to be prone to noise. So we apply a Gaussian smoothing filter to reduce noise.

3.2 Clustering by Hidden Markov Models

Humans perform gestures with variations in speed and position. To handle these variations, HMMs are used in this paper. An HMM is a probabilistic state machine and is widely used in recognition of dynamic processes. The Forward-Backward algorithm is an effective hill-climbing method for learning HMM parameters of observation sequences. And the Forward or Viterbi algorithm is used to evaluate the likelihood between observation and HMM[12].

HMMs provide a proper distance metric for sequence comparison. The distance between two sequences is computed as:

$$Dist(O_1,O_2)=\frac{1}{2}\left[\frac{1}{T_1}\big(P(O_1|\lambda_1)-P(O_1|\lambda_2)\big)+\frac{1}{T_2}\big(P(O_2|\lambda_2)-p(O_2|\lambda_1)\big)\right] \quad (1)$$

where λ_1,λ_2 denote two HMM models trained on sequences O_1, O_2; T_1,T_2 are the lengths of O_1, O_2, respectively.

The distance metric is used in [12] to compare HMM models. Considering the HMM as a generative model of sequences, the distance of models represent the distance of observations well.

Given a distance metric, many methods can be used to cluster the observation into several groups. In this paper, we choose hierarchical clustering to generate clusters from the observation[13]. The complete-link algorithm is used in this paper.

The operation of hierarchical clustering is a sequential process of merging the two most similar clusters to form a larger cluster. At the start, every sample is placed in its own cluster. The process is stopped when the distance between the two most similar clusters exceeds a threshold. In the complete-link algorithm, the distance between two clusters is maximum of all pairwise distances between samples in the two clusters. Compact clusters are produced by using the algorithm, and the result is fit for our purpose.

For N gesture segments, the whole clustering process requires training of N sequences, evaluation of N^2 distances, and a hierarchical clustering process on N samples. For large N, the process may be impractical. We can randomly select a set of segments to form the original clusters, use one HMM model to represent one cluster, and incrementally add the others.

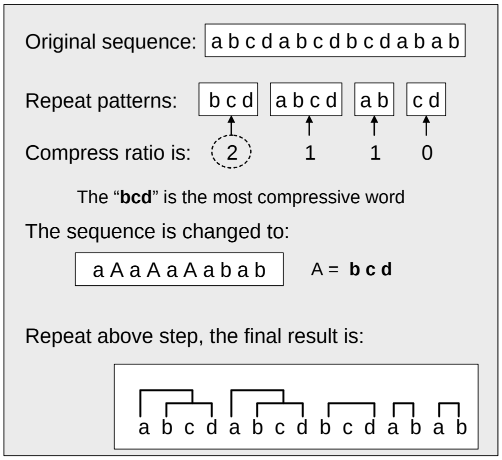
Original sequence:
a b c d a b c d b c d a b a b
Repeat patterns:
b c d
a b c d
a b
c d
Compress ratio is:
2
1
1
0
The “**bcd**” is the most compressive word
The sequence is changed to:
a A a A a A a b a b
A = **b c d**
Repeat above step, the final result is:
a b c d a b c d b c d a b a b

Fig. 2. An example of the COMPRESSIVE algorithm

3.3 Extracting Lexicon

Replacing each segment with its corresponding cluster number, we generate a discrete representation of the original continuous observation. Next we need to infer words from the sequence.

Extracting lexicon from a discrete sequence can be regarded as a language problem where we determine the basic vocabulary from a text. Considering the number of words, there may be a large number of possible solutions. We perhaps cannot find the true set of original base words, but such vocabulary has the property of providing a compact representation of the original data. So we can select an optimal solution by the MDL criteria, which is widely used in unsupervised learning and known to give interpretable results.

Directly finding the MDL solution is an NP-hard problem. Instead of finding the globally optimal solution, we adopt a heuristic approach called COMPRESSIVE [14]. It can be explained as selecting the word that provides the highest compression ratio of the input sequence. The compression ratio of a word is defined as:

$$\Delta DL = M \cdot N - (M + N + 1) \tag{2}$$

where M is the length of the word, and N is the number of repeated occurrences. This rule exhibits a tradeoff between pattern occurrence frequency and pattern length. In practice, the compressive first rule provides good performance for lexicon acquisition. An example is illustrated in Figure 2.

Our implementation uses a suffix-array [15]and has a complexity of $O(N^2)$. A suffix-array is a sorted list of all suffixes of a string, and can be constructed by initializing an array of pointers to every token in the string and sorting the array according to the lexicographic ordering of the suffixes denoted by the pointers.

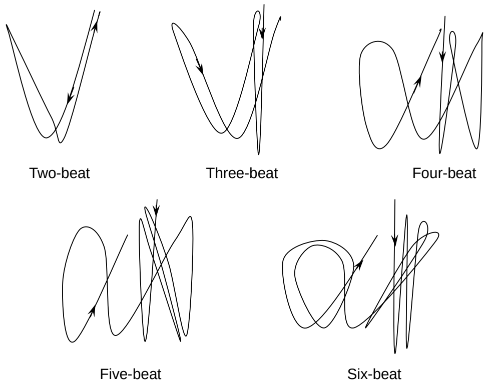

Fig. 3. Five basic conducting gestures.

To find the most compressed string, we need only scan the suffix array once and select one according to the Equation 2.

The lexicon obtained by our approach has a hierarchical structure. A word may contain other words. For example, in Figure 2, the string ($abcd$) has a sub-string (abc).

4 Experiment

To test our approach we carried out an experiment on musical conducting. It is natural to consider conducting as complex gestures consisting of a combination of simpler parts, for which we want to extract primitive patterns. We capture the data from a professional conductor who uses natural and precise conducting gestures.

We recorded about 8 minutes of conducting gestures. The whole sequence contains 5 basic beat-patterns, namely a two-beat pattern, three-beat pattern. Each basic pattern is performed many times. The prototype of each pattern is shown in Figure 3.

We use an optical motion capture system to obtain the 3d-position of the conductor's hand. The 3d-position is projected onto a virtual image plane, thus creating 2-d observation vector sequence contain the x, y position of conductor hand.

Totally 163 segments are obtained from temporal segmentation. We train a 5-state Gaussian HMM on each segment and calculate the distance matrix of sequences, which is shown in Figure 4. To choose the number of clusters, we draw a graph of the distance of the merged clusters. The cluster number can be estimated at a point, where increasing the number of clusters will merge very similar clusters. In our experiment the cluster number is selected as 15, as seen in Figure 4.

We obtain discrete representation of the original gesture sequence after the clustering process. The sequence is fed to the COMPRESSIVE algorithm to

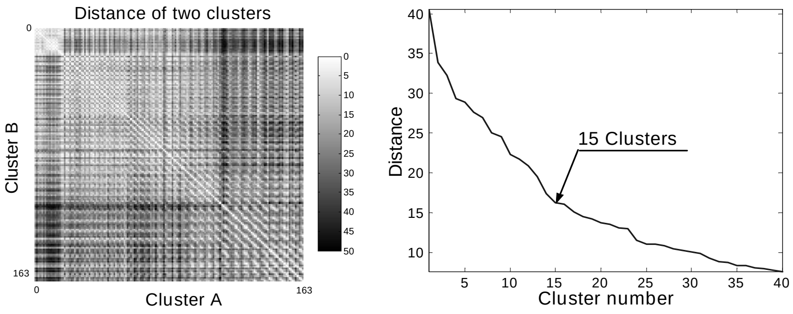

Fig. 4. Distance matrix of the sample(left), and the distance of merged clusters with respect to the number of clusters (right)

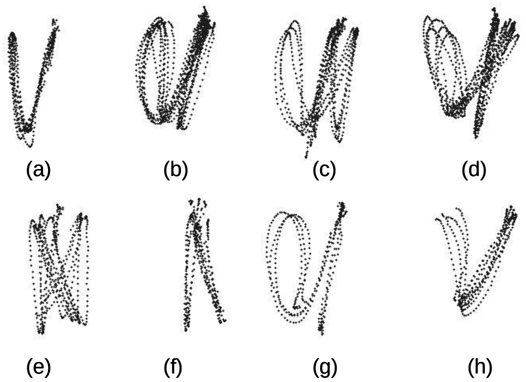

Fig. 5. The extracted patterns. (a) Two-beat. (b) Three-beat. (c) Four-beat. (d) Six-beat. Five-beat patterns are divided into (e) and part of (g). Within a hierarchical structure, (f) and (h) are part of (d) and (c),respectively.

determine words. Totally 8 words are extracted, including two-beat, three-beat, four-beat, and six-beat patterns. The five-beat pattern is not extracted because of few occurrences. The pattern is subdivided into two patterns as illustrated in Figure 5. The whole learning process takes approximate 4 minutes on a PIII 500MHZ CPU.

In the entire sequence, 104 of 163 segments are correctly labeled as prior known patterns. Although the rest are not recalled, they have been precisely assigned as new patterns.

It is shown that our algorithm can successfully segment and label continuous human gestures. Of course, this also reflects the limitation of the algorithm: it only extracts potential words instead of definite lexicon.

5 Conclusion

We have presented an approach for unsupervised clustering of human gestures. Our experiments show that our approach is feasible and useful. In addition, our approach can be easily adapted to incremental and online frameworks.

Our approach depends on the frequency of gestures. So it is suitable for structured human motion in which every gesture is repeated many times. Our future work will extend our approach to general human gestures. Another direction of interest is to apply our approach on partially labeled data.

References

1. D.M.Gavria. : The Visual Analysis of Human Movement: A Suvey. Computer vision and Image Understanding, Vol 73, 82-98,(1999).
2. T. Starner, A.Pentland.: Visual Recognition of American Language Using Hidden Markov Models. In Int workshop on Automatic Face and Gesture Recognition, 189-194, (1995).
3. Lee Campbell, Aaron.Bobick.: Recognition of Human Body Motion Using Phase Space Constraints, Fifth International Conference on Computer Vision, 624-630, Cambridge MA (1995)
4. Ying Wu, Thomas Huang.: Vision-Based Gesture Recognition: A Review. International Gesture Workshop, France, (1999)
5. Vladimir Pavlovic, James M.Rehg, John MacCormick.: Impact of Dynamic Model learning on Classification of Human Motion. International Conference on Computer Vision. (1999)
6. Brian.Clarkson, Alex.Pentland.: Unsupervised Clustering Of Ambulatory Audio and Video. AAAI99, (1999)
7. Matthew.Brand: Learning Concise Model of Human Activity from Ambient Video via a Structure-inducting M-step Estimator. MERL Technical report. (1997)
8. M. Walter, A.Psarrou, S. Gong.: An Incremental Learning Approach to Human Gesture Recognition Using Semi-CONditional DENSity PropagATION. International Conference on CARV, Singapore, (2000)
9. Nevil-Manning, and I. Witten.: Identifying Hierarchical Structure in Sequences: a Linear-time Algorithm. Artificial Intelligence Research, Vol 7, 66-82, (1997)
10. Wolff.J.G.: An Algorithm for the Segmentation of an Artificial Language analogue. British Journal of Psychology, vol 66, 79-90, (1975)
11. Kit. Chunyu.: A Goodness Measure for Phrase Learning via Compression with the MDL Principle. IESSLLI-98 Student Session, Chapter 13, 175-187, (1998).
12. L. Rabiner, B.Juang.: Fundamentals of Speech Recognition. Prentice Hall, New Jersey, USA (1993)
13. A.K.Jain, M.N.Murthy, P.J.Flynn.: Data Clustering: A Review. Technical report MSU-CSE-00-16, MSU, (2000).
14. Nevill-Manning, I. Witten.: Online and Offline Heuristics for Inferring Hierarchies of Repetitions in Sequence, Proceedings of the IEEE, in press.
15. K. Sadakane, H. Imai.: Constructing Suffix Arrays of Large Texts. Proc of DEWS98, (1998).