Auto clustering for unsupervised learning of atomic gesture components using minimum description length.

Michael Walter[1]
Alexandra Psarrou[1]
Shaogang Gong[2]

[1] Harrow School of Computer Science, University of Westminster
[2] Department of Computer Science, Queen Mary and Westfield College, London

Auto Clustering for Unsupervised Learning of Atomic Gesture Components using Minimum Description Length

Michael Walter † Alexandra Psarrou † and Shaogang Gong ‡
† Harrow School of Computer Science, University of Westminster,
Harrow HA1 3TP, U.K. zeoec,psarroa@wmin.ac.uk
‡ Dept. of Computer Science, Queen Mary and Westfield College,
London E1 4NS, U.K. sgg@dcs.qmw.ac.uk

Abstract

We present an approach to automatically segment and label a continuous observation sequence of hand gestures for a complete unsupervised model acquisition. The method is based on the assumption that gestures can be viewed as repetitive sequences of atomic components, similar to phonemes in speech, governed by a high level structure controlling the temporal sequence. We show that the generating process for the atomic components can be described in gesture space by a mixture of Gaussian, with each mixture component tied to one atomic behaviour. Mixture components are determined using a standard EM approach while the determination of the number of components is based on an information criteria, the Minimum Description Length.

Keywords: Gesture Recognition, Automatic Segmentation, Automatic Labelling, Data Driven Model Acquisition, Model Order Selection, Minimum Description Length (MDL), Atomic Gesture Components, Unsupervised Learning.

1. Introduction

Natural gestures are expressive body motions with underlying spatial and in particular temporal structure, which is probabilistic and often ambiguous. To account for such characteristics, the temporal structure of gestures can be modelled as stochastic processes under which salient phases of the structure are modelled as states and prior knowledge on both state distributions and observation covariances is learned from training examples [3, 15, 17, 11, 19]. However, the collection of training examples as well as the determination of states requires the segmentation and alignment of gestures. This task is ill-conditioned due to measurement noise, non-linear temporal scaling based on variations in speed and most notably human variation in the performing of a gesture. As a result the segmentation in gesture recognition typically involves manual intervention and hand labelling of image sequences.

In this paper we present a method to automatically segment and cluster continuous observation sequences of natural gestures for the unsupervised acquisition of gesture models, using only contextual information derived from the observation sequence itself. Our work is motivated by recent research in the field of Natural Gestures that identified two basic gesture types. Gestures based on two movement phases, away from a rest position into gesture space and back to the rest position and gestures based on three movement phases, away from the rest position into gesture space (preparation), followed by a small movement with hold (stroke) and back to the rest position (retraction) [12]. Our approach, therefore, is based on the assumption that gestures can be viewed as a recurrent sequence of atomic components, similar to phonemes in speech, starting and ending in rest positions and governed by a high level structure controlling the temporal sequence.

Automatic temporal gesture segmentation and partitioning into atomic components is achieved through a multi-scale analysis of the input trajectories for discontinuities, such as those that occur between preparation and stroke and rest positions that occur after strokes or transition into and out of gesture space. We show that atomic components, once normalised and projected into a gesture space, form clusters that correspond to atomic components. The density distribution of this gesture space can be described by a mixture of Gaussian, where each mixture component, k, models a different atomic component. Consequently, the determination of atomic components requires the determination of an optimum number of unknown clusters K, known as the problem of model order selection, and the estimation of the model parameters μ_k for $k = 1 \ldots K$.

Maximum likelihood methods such as k-means [8] or Expectation-Maximisation (EM) [5] provide effective tools for the determination of mixture components. However, the resulting mixture model depend on the a priori knowledge of the number of mixtures. The model order can be determined using constructive algorithms that employ cross validation techniques for model training [13], however the

0-7695-1074-4/01 $10.00 © 2001 IEEE

disadvantage of such methods is that they require a validation set, which in our case is not available. Alternative approaches to determine the number of clusters are based on information criteria, such as A Information Criterion (AIC) [1], Bayesian Information Criterion (BIC) [16] and Minimum Description Length (MDL) [14]. In the following sections we will show how MDL can be used to automatically segment the components within gesture space into clusters that correspond to atomic gestures, without any a priori knowledge on the number of atomic components present.

The rest of this paper is organised as follows. We first give an overview of related work. In Section 3 we describe the temporal segmentation of gestures and their partitioning into atomic components. In Section 4 we describe the gesture space representation and in Section 5 we show how MDL can be used for the automatic clustering of atomic components in gesture space. We describe experiments on the clustering in Section 6 and conclude in Section 7.

2 Related Work

The following paragraph gives a short review on papers addressing the concepts of automatic segmentation and unsupervised model acquisition.

Engel and Rubin [6] describe an approach for the qualitative classification of motion events. The events consist of smooth starts, smooth stops, pauses, impulse starts and impulse stops and are considered as motion events that partition a global motion into its psychological parts. Their method is based on a polar velocity representation and the derivation of first and second order derivatives. *Wilson et al* [21] presented an approach for the qualitative classification of natural gestures into either bi-or tri-phasic gesture. They identify plausible rest-state configurations of a speaker telling a story and parse the sequences in between into either bi-or tri-phasic gestures using a priory knowledge of the temporal structure describing both gestures types. *Wilson and Bobick* [20] describe an adaptive approach for unsupervised online learning of simple gestures for interactive control. Their algorithm requires a model of the temporal structure for the gesture to be learned, combined with contextual information derived from the application to bias the system in the early stages of runtime. *Vogler and Metaxas* [18] present an approach to continuous, whole-sentence American Sign Language (ASL) recognition, based on a sequential phonological model of ASL. They break ASL into movements and holds, both are considered phonemes and subsequently train Hidden Markov Models to recognise the phonemes, instead of whole signs. *Galata et al* [7] present an approach for the acquisition of statistical models of structured and semantically rich behaviour. Activities are modelled as sequence of atomic behaviour components, with variable length Markov models controlling the high level structure. Atomic behaviour components are seen as prototype sequence between two key prototypes, which in turn are identified as prototypes within the sequence, where changes drop below a preset threshold. The method can be used to generate and predict realistic human behaviour but can not generalise to previously unseen sequences. *Johnson* [10] presents an approach for the automatic acquisition of statistical behaviour models from continuous observations of long image sequences and derives a method for the assessment of behaviour typicality by exploiting the statistical nature of his behaviour models.

3 Temporal Segmentation

Temporal segmentation partitions a continuous observation sequence into plausible atomic components. Our approach is motivated by recent research in the field of Natural Gestures [12], that has identified five basic hand gesture types, iconic, metaphoric, cohesive, deistic and beat gestures. All gestures have their temporal signature in common. Gestures are typically embedded by the hands being in a rest state and can be divided into either bi-phasic or tri-phasic gestures. Beat and deictic gestures are examples for bi-phasic gestures. They have just two movement phases, away from the rest state into gesture space and back again, while iconic metaphoric and cohesive gestures have three, preparation, stroke and retraction. They are executed by transitioning from a rest state into gesture space (preparation), this is followed by a small movement with hold (stroke) and a movement back to the rest state (retraction).

The complete observation sequence, is recorded as a continuous sequence of $2D$ vertices, containing the x and y positions of a person's moving hand in an image plane. Segmentation is done in two steps. In a first step the complete observation sequence is analysed for segments where the velocity drops below a pre-set threshold to identify rest positions and pause positions that typically occur in bi-phasic gestures between transition into and out of gesture space and in tri-phasic gestures between stroke and retraction. A second step analyses the segments for discontinuities in orientation to recover strokes. The applied method is based on Asada and Brady's Curvature Primal Sketch [2], depicted in Figure 1.

4. Gesture Space Representation

Each atomic component extracted from the trajectory of a person's moving hand, consists of c 2D vertices $v_c = [x_1, y_1, x_2, y_2, ..., x_c, y_c]$ with each component having a different number of vertices c. Clustering algorithms, however work on d-dimensional sets of N input vectors $Z = [z_1, z_2, ..., z_N]$. This requires to transform the atomic components into a gesture space Figure 2. The transforma-

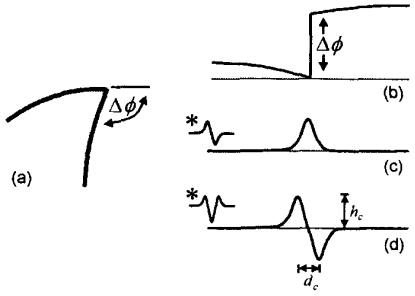

Figure 1: Discontinuity detection. The orientation of a two dimensional hand trajectory $f(t)$ is convoluted with the first N'_σ and second derivative N''_σ of a Gaussian $N_\sigma(t) = (1/(\sqrt{2\pi}\sigma)exp(-t^2/2\sigma^2)$ at different temporal scales $\sigma = \{\sigma_{min} \cdots \sigma_{max}\}$. The filter responses are analysed for characteristic maxima and zero crossings. Only discontinuities consistent over a large scale are registered, thus taking care of noise on different levels. a) Example trajectory containing a curvature discontinuity $\Delta\Phi$. b) The trajectory in orientation space, relating the orientation of the curve to the arc length along the curve. c) Filter response $N'_\sigma * f$ of the orientation of the trajectory $f(t)$ convoluted with the first derivative of a Gaussian $N_\sigma(t)$. d) Filter response $N''_\sigma * f$ of the orientation of the trajectory $f(t)$ convoluted with the second derivative of a Gaussian $N_\sigma(t)$. As shown in d) corners give rise to a pair of peaks with a separation $d_c \approx 2\sigma$ and height $h_c \approx |\Phi|/(\sqrt{2e\pi}\sigma^2)$. Note, d_c is linearly dependent on the scale constant σ and monotonically decreases with σ, which provides a strong clue for the detection of corners.

tion consists of three steps. First the number of 2D vertices is normalised. The atomic components are approximated by splines, interpolated into d vertices and stored as d-dimensional vector $z_i = [x_1, y_1, x_2, y_2, ..., x_d, y_d]$. In a second step each vector z_i is concatenated with a scale factor $s_i = (c_i - c_{min})/(c_{max} - c_{min})$, the ratio of the original number of vertices minus the minimal number and the maximal number minus the minimal number of vertices to preserve information on the original length. Finally redundant dimensions are removed using Principal Component Analysis (PCA). Figure 3 shows an example for the weight distribution in gesture space used to represent the normalised 41-dimensional atomic components.

The atomic components projected into gesture space form clusters, that can be approximate by a mixture of K Gaussian, defined by mixture coefficients w_k, d-dimensional means μ_k and Covariances Σ_k, with each mixture component corresponding to one atomic behaviour.

$$f(z_i|W_K, \Theta_K) = \sum_{k=1}^{K} w_k N(z_i, \mu_k, \Sigma_k) \qquad (1)$$

This allows us to equate the determination of the mixture components with the determination of the atomic components itself. There is a considerable amount of literature on the estimation of mixture parameter and standard Expectation Maximisation methods can be used to determine the values of Σ_k, μ_k and w_k for a known model order K. There are no methods to determine the number K of parameter directly, however iterative procedures based on information criteria can be used as described in the next section.

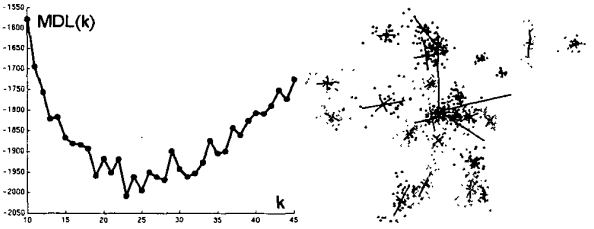

Figure 2: The Minimum Description Length MDL(k) calculated for each cluster configuration [$10 < k < 45$] of the atomic components, with global minimum determined for 23 cluster (left). Gesture Space segmented into 23 cluster: The Projection of the 3 largest Principle Components of each atomic component (right).

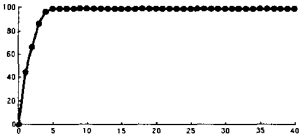

Figure 3: The weight distribution in gesture space used to represent the normalised 41-dimensional atomic components. 95% of the original information is contained in the first 5 Principle Components.

5. Extracting Atomic Components

The problem of model order selection has been widely studied in literature (see [4] for a review). Heuristic methods have been proposed by Akaike [1], Schwarz [16] and Rissanen [14] [(AIC) A Information Criterion, (BIC) Bayesian Information Criterion, (MDL) Minimum Description Length]. They are heuristic in the sense that they do not minimise an error function between an estimated and the true model order. Instead they define various information criteria that only depend on the unknown model order K, which is defined as minimising value for the respective criterion. One of the most popular criteria, the information criteria of Rissanen MDL is defined as

$$MDL(K) = -ln[L(Z|W_K, \Theta_K)] + \frac{1}{2}Mln(N) \qquad (2)$$

MDL is obtained from information-theoretic considerations, and the model order is defined as the model that minimises the description length, i.e. the model that encodes the vector of observations in the most efficient way [9].

The first term $-ln[L(Z|W_K, \Theta_K)]$, the maximised mixture likelihood of $P(Z|W_K, \Theta_K)$, measures the systems entropy and can be seen according to Shannon's Information Theorem as a measure for the number of bits needed to encode the observations $Z = [z_1, z_2, ...z_N]$, with respect to the model parameter W_K and Θ_K

$$P(Z|W_K, \Theta_K) = \prod_{i=1}^{N} f(z_i|W_K, \Theta_K) \quad (3)$$

The second term, $\frac{1}{2}Mln(N)$ measures the additional number of bits needed to encode the model parameter and serves as penalty for models that are too complex. M describes the number of free parameter and is given for a Gaussian mixture by $M = 2dK + (K - 1)$ for (K-1) adjustable weights due to the constraint $\sum_k w_k = 1$ and $2d$ parameter for d dimensional means and diagonal Covariance matrixes.

The optimal number of clusters and therefore number of atomic components can be determined by applying the following iterative procedure.

1. For all K, $\{K_{min} < K < K_{max}\}$

   (a) Maximise the likelihood $L(Z|W_k, \Theta_k)$ using the k-means [8] or iterative EM [5].

   (b) Calculate the value of MDL(k) according to Equ. (2) and Equ. (3).

2. Select the model parameters $\{W_k, \Theta_k\}$ for the minimising value of MDL(k).

6. Experiments

To evaluate our approach, we recorded a participant performing 7 gestures in arbitrary order. The recording included deictic gestures such as "pointing left" and "pointing right", metaphoric gestures such as "he bent a tree" and "there was a big explosion" and communicative gestures such "waving high", "waving low" and "please sit down". Examples are shown in Figure 5, Figure 6 and Figure 7.

Figure 4: Experiminetal Setup: The Polhemus tracker with sensors attached to the right hand and head (left). The recorded input sequence segmented into 668 atomic components (right).

A continuous sequence of gestures was recorded in a time window of 10 minutes. Gestures were performed

Figure 5: Deictic gestures: *"pointing left"* (top 2 rows), *"pointing right"* (bottom 2 rows)

Figure 6: Metaphoric gestures: *"he bent a tree"* (top 2 rows), *"there was a big explosion"* (bottom 2 rows)

roughly 20 times in random order. The gestures were recorded using a Polhemus tracker, an electromagnetic tracking device that is able to determine the 3D position of a small sensor relative to the center of a transmitter. The experimental set-up is shown Figure 4 (left). Two sensors were attached to the participant's body, one to the head, used for reference and one to the right hand. The 3D positions of both sensors were recorded with 5 frames per second and the relative difference projected onto a virtual image plane, thus creating a 2-dimensional observation trajectory containing the [x,y] position of the participant's hand relative to the head. The recorded input sequence was automatically partitioned into 668 atomic components (see Figure 4 (right)).

All atomic components were transformed into "Gesture

Figure 7: Communicative gestures: *"waving high"* (top 2 rows), *"waving low"* (middle 2 rows) and *"please sit down"* (bottom 2 rows)

Space": each component was converted into a spline, interpolated into 20 $2D$ vertices and stored as 41-dimensional vector (20 $2D$ vertices plus scale factor) and reduced to 5 dimensions using PCA, still containing 95% of the original information (Figure 3).

The resulting distribution was approximated by a mixture of Gaussian with $[10 < k < 45]$ mixture components, determined using a standard k-means clustering algorithm. The Minimum Description Length MDL(k) was computed for each configuration and a global minimum was determined for $k = 23$ mixtures (Figure 2 (left)).

Figure 8, Figure 9 and Figure 10 show examples of gestures (leftmost columns) and their associated atomic components. The small number next to each component indicates the mixture identifier. The small squares at the end of each atomic component trajectory indicate the direction of the primitive movement and are there to aid visual understanding of the components.

Figure 8 shows two bi-phasic gestures, *"pointing left"* and *"pointing right"* . The leftmost column shows the complete movement of each of the two gestures: Away from the rest state into gesture space and back again, whereas the middle and right columns show the atomic components extracted for each of the gestures. The middle column shows the atomic components (or primitive movements) into gesture space, and the right column shows the atomic components out of gesture space.

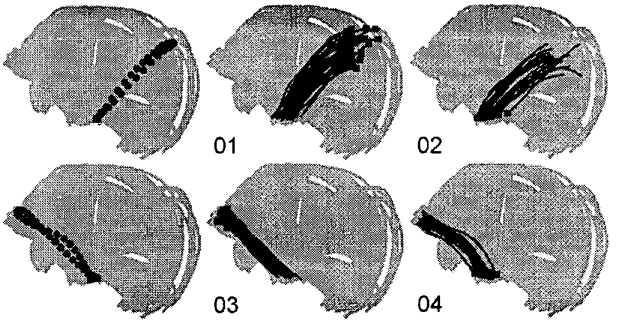

Figure 8: Left column: Example trajectories of gestures *pointing right*(top) and *pointing left*(bottom). Middle column: Corresponding atomic components into gesture space. Right column: Corresponding atomic components out of gesture space.

Figure 9 shows examples for gestures that can be segmented into three atomic components. From top to bottom they are: *"please sit down"*, *"he bent a tree"*and *"there was a big explosion"* . The leftmost column shows the complete movement of each of the three gestures, whereas the remaining columns show the extracted atomic components corresponding to each gesture.

Some of the atomic components are shared by two or more gestures. The atomic component represented by the cluster identified as "07" is shared by both *"please sit down"* and *"there was a big explosion"* gestures. Similarly the atomic component represented by the cluster identified as "01" is shared by both *"pointing left"* (Figure 8), and *"there was a big explosion"* gestures. Finally Figure 10 shows examples of gestures *"waving high"* and *"waving low"* , that can be segmented into four atomic components. The atomic component represented by the cluster identified as "08" is shared by both *"he bent a tree"* (Figure 9), and *"waving high"* gestures.

In total we extracted 23 clusters. 18 clusters are associated with atomic components corresponding to primitive gesture movements. Each of these 18 clusters have a population of 20-40 components each. The remaining 5 unassigned clusters shown in Figure 11 can not directly be related to any gestures. They have a population of 5-10 short atomic components and are the result of segmentation artefacts.

7. Summary and Conclusions

We presented an approach to automatically segment and label a continuous observation sequence into atomic components, using only contextual information derived from the observation sequence itself. We assumed that gestures can be viewed as a repetitive sequence of atomic components

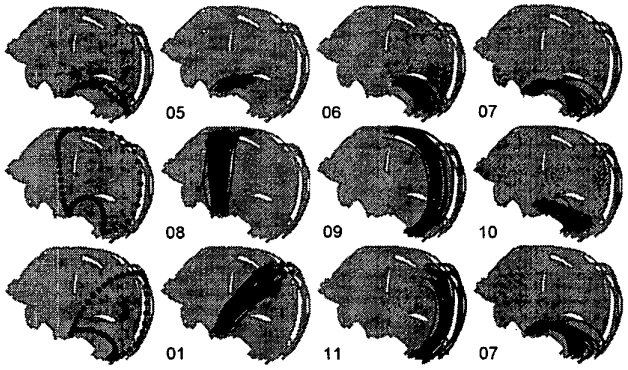

Figure 9: Gestures that can be partitioned in three atomic components. Leftmost column: Example trajectories of gestures *please sit down* (top row), gestures *he bent a tree* (middle row) and *there was a big explosion* (bottom row). Remaining columns: Corresponding atomic components into and out of gesture space.

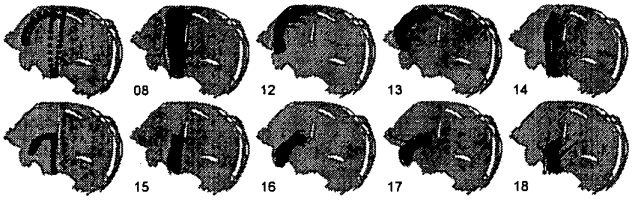

Figure 10: Gestures that can be partitioned into four atomic components. Leftmost column: Example trajectories of gestures *waving high* (top) and *waving low* (bottom). Remaining columns: Corresponding atomic components into and out of gesture space.

Figure 11: The remaining 5 clusteres are based on segmentation artefacts

that can be modelled in a gesture space by mixture of Gaussian. Mixture components were determined using a standard k-means clustering algorithm and the number of components was automatically determined using the MDL criterion. Visual inspection suggests that the determined number is equal to the real number of underlying components. However, a common criticism of k-means and related maximum likelihood methods is that they do not address the problem of noise and assign all input elements to one particular class. Therefore we will look into more 'loosely' defined methods able to model noise. In the future we will address the higher level structure controlling the temporal sequence and consequently will work on a model describing the sequence of atomic components.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] H. Asada and M. Brady. The curvature primal sketch. *Technical Report. MIT AI memo 758*, 1984.

[3] A. Bobick and A. Wilson. A state-based technique for the summarisation of recognition of gesture. *ICCV*, pages 382–388, 1995.

[4] H. Bock. Probability models and hypothesis testing in partitioning cluster analysis. *Clustering and Classification*, pages 377–453, 1996.

[5] N. Dempster, A.P.and Laird and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:185–197, 1977.

[6] S. Engel and J. Rubin. Detecting visual motion boundaries. *Workshop on Motion: Representation and Analysis*, pages 107–111, 1986.

[7] A. Galata, N. Johnson, and D. Hogg. Learning behaviour models of human activities. *BMVC*, pages 12–22, 1999.

[8] J. Hartigan and M. Wong. A k-means clustering algorithm. *Appl. Statist.*, 28:100–108, 1979.

[9] D. Hirshberg and N. Merhav. Robust methods for model order estimation. *IEEE Transactions on Signal Processing*, 44:620–628, 1996.

[10] N. Johnson. Learning object behaviour models. *PhD thesis, University of Leeds*, England, September 1998.

[11] S. McKenna and S. Gong. Gesture recognition for visually mediated interaction using probabilistic event trajectories. *BMVC*, pages 498–508, Southampton, England, September 1998.

[12] D. McNeill. Hand and mind: What gestures reveal about thought. *University of Chicago Press*, 1992.

[13] Y. Raja, S. McKenna, and S. Gong. Colour model selection and adaptation in dynamic scenes. *ECCV*, Freiburg, Germany, 1998.

[14] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[15] J. Schlenzig, E. Hunter, and R. Jain. Recursive identification of gesture inputs using hidden markov model. *Workshop on Applications of Computer Vision*, pages 187–194, 1994.

[16] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[17] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. *International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, Zurich, 1995.

[18] C. Vogler and D. Metaxas. Toward scalability in asl recognition: Breaking down signs into phonemes. *Gesture Workshop*, Gif sur Yvette, France 1999.

[19] M. Walter, S. Gong, and A. Psarrou. Stochastic temporal models of human activities. *International Workshop on Modeling People*, pages 87–94, Corfu, Greece, 1999.

[20] A. Wilson and A. Bobick. Realtime online adaptive gesture recognition. *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real Time Systems*, 1999.

[21] A. Wilson, A. Bobick, and J. Cassell. Temporal classification of natural gesture and application to video coding. *IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, pages 948–954, 1997.