Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Finding Motion Primitives in Human Body Gestures

Reng, Lars; Moeslund, Thomas B.; Granum, Erik

Published in:
GW 2005

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Reng, L., Moeslund, T. B., & Granum, E. (2006). Finding Motion Primitives in Human Body Gestures. In S. Gibet, N. Courty, & J. F. Kamp (Eds.), *GW 2005* (pp. 133-144). IEEE Computer Society Press. Lecture Notes In Artificial Intelligence No. 3881

Finding Motion Primitives in Human Body Gestures

Lars Reng, Thomas B. Moeslund, and Erik Granum

Laboratory of Computer Vision and Media Technology,
Aalborg University, Denmark

Abstract. In the last decade speech processing has been applied in commercially available products. One of the key reasons for its success is the identification and use of an underlying set of generic symbols (phonemes) constituting all speech. In this work we follow the same approach, but for the problem of human body gestures. That is, the topic of this paper is how to define a framework for automatically finding primitives for human body gestures. This is done by considering a gesture as a trajectory and then searching for points where the density of the training data is high. The trajectories are re-sampled to enable a direct comparison between the samples of each trajectory, and enable time invariant comparisons. This work demonstrates and tests the primitive's ability to reconstruct sampled trajectories. Promising test results are shown for samples from different test persons performing gestures from a small one armed gesture set.

1 Introduction

In the last decade speech synthesis and speech recognition have transferred from only being research topics into core technologies in commercially available products. One of the key reasons for this transfer is the identification and use of an underlying set of generic symbols constituting all speech, the phonemes. Phonemes are basically small sound samples that put together in the correct order can generate all the words in a particular language, for example English.

It is widely accepted that more than half of the information transmitted in a human-human interaction is done by other means than speech, and that the human body language is responsible for most of this information. Furthermore, for better human-computer interfaces to be build the computer might need to be equipped with the ability to understand the human body language [15]. Since automatic recognition of human body language is a desired ability research has been conducted in this area. Much of this research is based on defining a subset of the human body language, normally denoted "actions", and then building a classifier based on some kind of learning scheme applied to some training data. The result of the training is a sequence of values in some state-space for each action. The different learnt sequences are compared to the input data during run-time and a classification is carried out.

In some systems, however, a different approach is followed[1]. This approach is based on the idea that an action can be represented by a set of shorter (in terms of time duration) primitives. These primitives take different names such as movemes [4], atomic

---

[1] These approaches are sometimes motivated directly by the notion of finding "phonemes" in the human body language.

movements [5], activities [2], behaviors [12, 17], snippets [9], dynamic instants [16], states [3], and examplars [14].

Besides the different names used to describe the notion of motion primitives, the approaches also differ in another way, namely whether a primitive is dependent or independent on time. The approaches based on independence find their inspiration in key-frame animation. Key-frame animation is based on the idea that animating an articulated object in a time sequence is a matter of defining the configurations for a number of distinct frames (key-frames) and then interpolate all in-between frames using e.g., inverse kinematics. Mapping this concept to the problem of recognizing human body language converts the problem to a matter of recognizing a number of single configurations and ignoring all in-between configurations. This concept is sound but introduces a number of problems including the problem of defining which configurations (or key-frames) that best represent an action.

In the work by Rao *et al.* [16] the problem of recognizing dynamic hand gestures is addressed. They track a hand over time and hereby generate a trajectory in 3D space (x- and y-position, and time). They search the trajectory for significant changes, denoted dynamic instants, which are defined as instants with a high curvature. In the work by Jordi [8] the problem of finding key-frames for cyclic actions, like walking and running, is addressed. They capture the joint angles using an optical motion capture system and compactly represent a time sequence of such data using a point distribution model. Since the actions are cyclic they argue that the likelihood of a configuration being part of an action can be measured as the Mahalanobis distance to the mean. The key-frames are then defined as configurations where the Mahalanobis distance locally is maximum, i.e., key-frames are the least likely configurations!

The alternative to the key-frame approach is to represent the entire trajectory (one action), but doing so using a number of smaller sub-trajectories. That is, the entire trajectory through a state space is represented as opposed to only representing a number of single points. Several problems are associated with this approach, for example, how to define the length of the sub-trajectories. If too long then the primitives will not be generic. If too short the compactness of the representation is lost. In the work by Hodgins *et al.* [7] different approaches to find such sub-trajectories for full body motion are compared, and show promising results. Their comparison of three different approaches finds Probabilistic PCA as a very efficient tool for finding transitions between different behaviours.

In the work by Howe *et al.* [9] the problem of capturing the 3D motion of a human using only one camera is addressed. The main body parts are tracked in 2D and compared to learned motion patterns in order to handle the inherent ambiguities when inferring 3D configurations from 2D data. The learned motion patterns are denoted "snippets" and consist of 11 consecutive configurations. These are learned by grouping similar motion patterns in the training data. In the work by Bettinger *et al.* [1] the problem of modeling how the appearance of a face changes over time is addressed. They use an active appearance model to represent the shape and texture of a face, i.e., one point in their state-space corresponds to one instant of the shape and texture. They record and annotate a number of sequences containing facial changes. Each sequence corresponds to a trajectory in their state space. The states with the high-

est densities are found and used to divide the data into sub-trajectories. These sub-trajectories are modeled by Gaussian distributions each corresponding to a temporal primitive.

The different approaches found in the literature that uses the notion of motion primitives more or lees follow the structure below.

Temporal content. Either only a single time instant define a primitive or a primitive is based on a consecutive number of temporal instants.

Motion capture. In order to find the primitives the motion data needs to be captured. This could for example be done by an optical system or electromagnetic sensors.

Data representation. What is measured by the motion capture system is normally the 3D position of the different body parts. These measurements are often represented used normalized angles. Furthermore, the velocity and acceleration might also be considered.

Preprocessing. The captured data can have a very high dimensionality and can therefore be represented more compactly using, e.g., PCA. Furthermore, the data might be noisy and is therefore often filtered before further processing.

Primitives. It needs to be decided how to define a primitive. Often this is done via a criteria function which local minima/maxima defines the primitives.

Application. The chosen method needs to be evaluated. This can be with respect to the number of primitives versus the recognition rate, but it can also be a comparison between the original data and data synthesized using the primitives.

Our long term goal is to find a set of generic primitives that will enable us to describe all (meaningful) gestures conducted by the upper body of a human. Our approach is to investigate different data representations together with different criteria functions. We seek to find primitives for both recognition and synthesis, and evaluate the relationship between the two.

This particular paper presents the initial work towards our goal and the focus of the paper is to obtain experiences with all the topics listed above. Concretely we define a number of one-armed gestures and for each gesture we evaluate a method used to find primitives. The criteria function is based on a combination of two concepts, namely the curvature and density of a trajectory.

The paper is structured as follows. In section 2 the gesture data and the applied motion capture technique are presented. In section 3 we describe how the data is normalized. In section 4 the concept behind the primitives is given. In section 5 we present the density measure used in the criteria function, and in section 6 we combine this with a distance measure and defined how the criteria function is evaluated in order to select the primitives. In section 7 the test results are presented and in section 8 a conclusion is given.

2   The Gesture Data

The gestures we are working with are inspired by the work of [13] where a set of hand gestures are defined. The gestures in [13] are primarily two-hand gestures, but we simplify the setup to one-hand gestures in order to minimize the complexity and focus
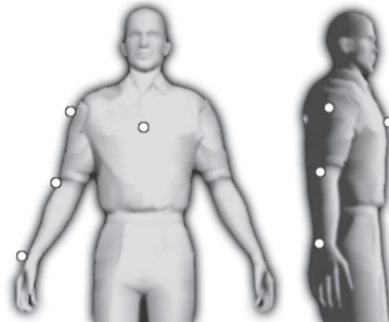
Fig. 1. Placement of sensors. The figure is adapted from [11].

on the primitives. Some of the gestures were exchanged with other more constructive ones. The final set of gestures are, as a result of this, all command gestures which can be conducted by the use of only one arm. The gestures are listed below.

Stop: Hand is moved up in front of the shoulder, and then forward (with a blocking attitude), and then lowered down.

Point forward: A stretched arm is raised to a horizontal position pointing forward, and then lowered down.

Point right: A stretched arm is raised to a horizontal position pointing right, and then lowered down.

Move closer: A stretched arm is raised to a horizontal position pointing forward while the palm is pointing upwards. The hand is then drawn to the chest, and lowered down.

Move away: Hand is moved up in front of the shoulder while elbow is lifted high, and the hand is then moved forward while pointing down. The arm is then lowered down.

Move right: Right hand is moved up in front of the left shoulder. the arm is then stretched while moved all the way to the right, and then lowered down.

Move left: Same movement as *Move right* but backwards.

Raise hand: Hand raised to a position high over the head, and then lowered down.

Each gesture is carried out a number of times by a number of different subjects, in order to have both data for inter-person comparisons, and comparable data for each gesture by several different subjects.

The gestures are captured using a magnetic tracking system with four sensors: one at the wrist, one at the elbow, one at the shoulder, and one at the torso (for reference), as shown in figure 1. The hardware used is the Polhemus FastTrac [10] which gives a maximum sampling rate of $25Hz$, when using all four sensors. In order to normalize the data and make it invariant to body size, all the collected 3-dimensional position data is converted to a time sequence of four Euler angles: three at the shoulder and one at the elbow. Besides normalizing the data, this transformation also decreases the dimensionality of the data from 12 to only 4 dimensions.

3   Normalizing the Data

In order to compare the different sequences they each need to be normalized. The goal is to normalize all the gesture trajectories so each position on a trajectory can be described by one variable t, where $t \in [0; 1]$.

The first step is to determine approximately where the gestures' endpoints are. In this experiment we have chosen to do so by defining a gesture set where all gestures are considered to both start and stop when the arm is hanging relaxed from the shoulder. A velocity threshold ensures that the small movements done between gestures is added to neither, and simplifies the separation of the individual gestures.

The trajectories are therefore homogeneously re-sampled in order to enable time invariant comparisons. This is done by interpolating each gesture, in the 4D Euler-space, by use of a standard cubic spline function. The time and velocity information is, however, still available from parameters in the new sample points, even though this is not used in this work. The homogeneously re-sampling allows for a calculation of the statistics for each gesture *and* at each sample point. Concretely, for each gesture we calculate the mean and covariance for each sample point, i.e., each instant of t. This gives the average trajectory for one gesture along with the uncertainties along the trajectory represented by a series of covariant matrices, see figure 2.
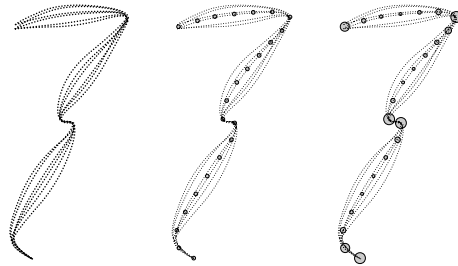
Fig. 2. Six example trajectories for a fictive gesture. Left: Input after cubic spline interpolation. Middle: Input including the position of the mean points. Right: The sizes of the mean points indicate the density of the curves.

4   Defining Primitives of Human Gestures

This section gives an intuitive description of which criteria define a good primitive candidate. In order to find the primitives we apply the following reasoning. A primitive is a particular configuration of the arm, i.e., of the four Euler angles. For a configuration to qualify as a good primitive candidate the configuration must appear in all the training data, at approximately the same time. For such a configuration to exist, all the training data must vary very little at this point in space and time, which will result in a very high density of training trajectories at this position in space. The density of a particular configuration measures how close the original sequences passed this configuration. The closer they passed the higher the density, which corresponds to

a good candidate. The logic behind this is very simple; only at the points where we have selected a primitive can we be sure that our new interpolated curve will parse directly though. Even though this paper does not go into detail with the recognition part, the main reasons for selecting primitives where the density is high is, that it makes good sense to compare an unknown curve to our known interpolated curve, at exactly the points where all the training data trajectories laid closest, see figure 2. However, just selecting the *n* points with the highest density will result in very inefficient primitives, since one primitive is enough to direct the interpolated curve through this area. So selecting primitives in places where the curve already passes by, will offer little to the reconstruction of the original curve. In the next two sections we describe how we calculate the density measure, and how this is used to select our primitives.

5   Measuring the Density

In section 3 the points constituting each trajectory were normalized so that the trajectories for different test subjects can be compared. That is, each trajectory was re-sampled so that they each consist of the same amount of points which are aligned. We can therefore calculate the covariance matrix for each time instant. The covariance matrices for each time instant express both how data are correlated but also how they are spread out with respect to the mean. The Mahalanobis distance expresses this relationship by defining a distance in terms of variances from a data point to the mean. It is defined as

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{1}$$

where $\mathbf{x}$ is a data point, $\boldsymbol{\mu}$ is the mean for this particular time instant, and $\mathbf{C}$ is the covariance matrix. If r is constant then equation 1 becomes a hyper ellipsoid in 4D space. The data points on its surface have the same variance-distance to the mean. The volume of a hyper ellipsoid with fixed Mahalanobis distance is a direct measure of the density of the data at this time instant. A big volume corresponds to a low density where the points are spread out, whereas a small volume corresponds to a high density as the same amount of data are located at a much smaller space. The volume of a hyper ellipsoid which is expressed as in equation 1 is given as [6]

$$V = \frac{\pi^2 \cdot r^4}{2} |\mathbf{C}|^{\frac{1}{2}} \tag{2}$$

where $|\mathbf{C}|$ is the determinant of the covariance matrix. We are not interested in the actual value of the volume but rather the relative volume with respect to the other time instants. Therefore equation 2 can be reduced to $V = |\mathbf{C}|^{\frac{1}{2}}$ and is illustrated in figure 2. Below we give an intuitive interpretation of this measure.

5.1   Geometrical Interpretation

Due to the inherent difficulty of illustrating in 4D we give the geometric interpretation of $|\mathbf{C}|^{\frac{1}{2}}$ in 2D and then generalize to higher dimensions.
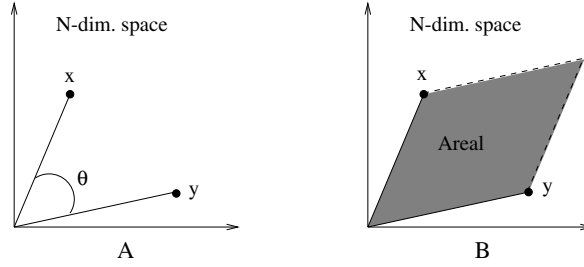
Imagine that we have N samples in the 2D X-Y plan. For simplicity we assume that the mean of the data is the origin. The covariance matrix is given as

$$\mathbf{C} = \begin{bmatrix} c_x^2 & c_{xy}^2 \\ c_{yx}^2 & c_y^2 \end{bmatrix} \tag{3}$$

where entries are calculated as

$$c_x^2 = \frac{\Sigma x^2}{n-1} \qquad c_y^2 = \frac{\Sigma y^2}{n-1} \qquad c_{xy}^2 = c_{yx}^2 = \frac{\Sigma xy}{n-1} \tag{4}$$

We can think of $\{x_1, x_2, x_3, \cdots, x_n\}$ and $\{y_1, y_2, y_3, \cdots, y_n\}$ as two vectors in an N-dimensional space, see figure 3.

Fig. 3. A: An illustration of the 2D data interpreted as vectors in an N-dimensional space. B: The area of the parallelogram spanned by the two vectors.

The angle, θ, between the two vectors expresses the correlation between the two variables, x and y. The smaller the angle the higher correlation. The correlation is often normalized to the interval $[0, 1]$ using trigonometry, i.e., the correlation is given as $\cos(\theta)$. Relating this to the two vectors yields

$$\cos(\theta) = \frac{\mathbf{x}^T\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{\Sigma xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} \tag{5}$$

Inserting this into the covariances yields

$$c_{xy}^2 = c_{yx}^2 = \frac{\sqrt{\sum x^2}\sqrt{\sum y^2}\cos(\theta)}{n-1} \tag{6}$$

We can now calculate the determinant of the covariance matrix as

$$|\mathbf{C}| = c_x^2 \cdot c_y^2 - c_{xy}^2 \cdot c_{yx}^2 \Rightarrow \tag{7}$$

$$|\mathbf{C}| = \frac{\sum x^2 \sum y^2}{(n-1)^2} - \left(\frac{1}{n-1}\right)^2 \left(\sqrt{\sum x^2}\right)^2 \left(\sqrt{\sum y^2}\right)^2 \cos(\theta)^2 \Rightarrow \tag{8}$$

$$|\mathbf{C}| = \left(\frac{1}{n-1}\right)^2 \sum x^2 \sum y^2 \left(1 - \cos(\theta)^2\right) \Rightarrow \tag{9}$$

$$|\mathbf{C}| = c_x^2 \cdot c_y^2 \cdot \sin(\theta)^2 = (c_x \cdot c_y \cdot \sin(\theta))^2 \tag{10}$$

This equation can be interpreted in a geometrical manner as illustrated in figure 3.B. The area, A, of the parallelogram (shaded area) is given as $A = c_x c_y \sin(\theta)$, i.e., the area depends on the standard deviation of the two variables, x and y, and the correlation between them. The higher the correlation the smaller the area. The 2D interpretation of $|\mathbf{C}|^{\frac{1}{2}}$ is the area of the parallelogram in the fourth potent. In the 3D case the geometric interpretation of the determinant of the covariance matrix is the volume of the parallelepiped spanned by three variables and their correlation. In 3D+ the geometric interpretation becomes less intuitive and is sometimes expressed as the generalization of the concept of variance.

6   Selecting the Primitives

Above we have defined and presented a method for calculating the density measure, and are now ready to include this into one criteria function that can be evaluated in order to find the primitives. The criteria function will combine the density measure with the distance between the homogeneously re-sampled mean gesture trajectory (m) and a trajectory made by interpolating the endpoints and the first selected primitives, using a standard cubic spline function (c) for each of the four Euler angles. In order to make a direct comparison, both the mean gesture trajectory and the interpolated cubic spline trajectory were given the same amount of points. This enables a calculation of the *error*-distance (δ) between the curves for each point pair. If multiplying this error distance at each point with the density (β), we can get a distance measure much similar to the Mahalanobis.

Since the four angles might not have the same dynamic ranges and more freedom to optimize future parameters is desired, the criteria function (λ) is defined as a weighted sum of error measures (α_i) for each of the four Euler angles:

$$\lambda(t) = \omega_1 \alpha_1(t) + \omega_2 \alpha_2(t) + \omega_3 \alpha_3(t) + \omega_4 \alpha_4(t) \qquad (11)$$

where the four weights $\omega_1 + \omega_2 + \omega_3 + \omega_4 = 1$, and the error measure:

$$\alpha_i(t) = \beta_i(t) \cdot \delta_i(t)^2, and : \delta_i(t) = \sqrt{(m_i(t) - c_i(t))^2} \qquad (12)$$

Given the criteria function in equation 11 we are now faced with the problem of finding the N best primitives for a given trajectory. The most dominant primitive, χ_1 is obviously defined as : $\chi_1 = arg \max_t \lambda(t)$.

In order to find the second primitive, the first one is added to the cubic spline function (c), and the four trajectories are then recalculated, so new error distance measures can be calculated, see figure 4. This procedure can be repeated until the sum of all (λ) falls below a given threshold, or the number of primitives reaches an upper threshold.

6.1   Optimizing the Primitive's Position

Knowing that this method can, most likely, be improved; we tried to implement an optimizing step at the end of each primitive selection. A brute force test on all the test data could be used in order to find the optimal solution given a number of maximum
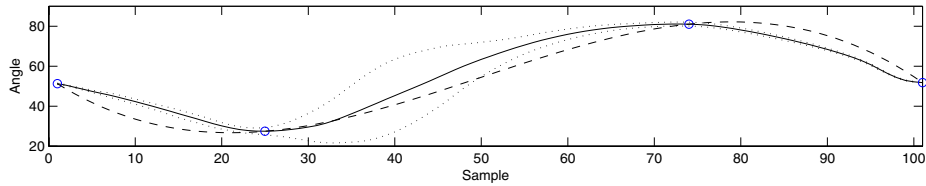
Fig. 4. Calculating the error-distance in one direction. Solid: The mean gesture trajectory. Dashed: Interpolated cubic spline. Dotted: Variance of training data. Circles: Selected primitives and endpoints.

primitives and number of samples. This is, however, very time consuming, and only valuable for the given data set, and was therefore not considered.

Instead, tests were done with another much faster method. After each new primitive was selected, all the selected primitives were tested in a position one step to each side along the mean gesture trajectory. Only if they could lower the total error sum, will they move to this position, and as long as just one primitive could be moved, all other would be tested again. This method will bring the error sum to a local minimum, but not to a guarantied global minimum.

See the following section for tests results on both previous described methods.

7   Results

The tests described in this section were made on a training data set based on the eight one arm gestures described in section 2. Three tests persons conducted each gesture no less than ten times resulting in a total of 240 gestures[2]. The evaluation of our approach consists of two tests for each action:

- Investigate how many primitives are required in order to reconstruct the original gestures.
- Evaluate the optimization step, and determine whether or not this should be used in our continuous work.

It is our belief that the only reasonable way to evaluate whether the reconstruction of a gesture is life like enough to look natural, is to have a robot or virtual human avatar performing the reconstructed gestures before a large number of test persons, and having these evaluate the result. This was however not within range of our possibilities at this point in our research. Instead, all reconstructions were evaluated by the research group from a large number of graphs such as those shown in figures 5 and 6. The graphs show the four angle spaces and error measure of the gesture *Move Left*, with two endpoints and 2,4 and 8 primitives. Figure 5 show the result of the reconstruction without the optimizing step, where as 6 were depicture the reconstruction of the exact same angle spaces, but with the optimization. The sum of the error measures for each curve pair of

---

[2] Additional 160 training gestures were made but had to be removed from the set do to extremely low signal to noise ratio.
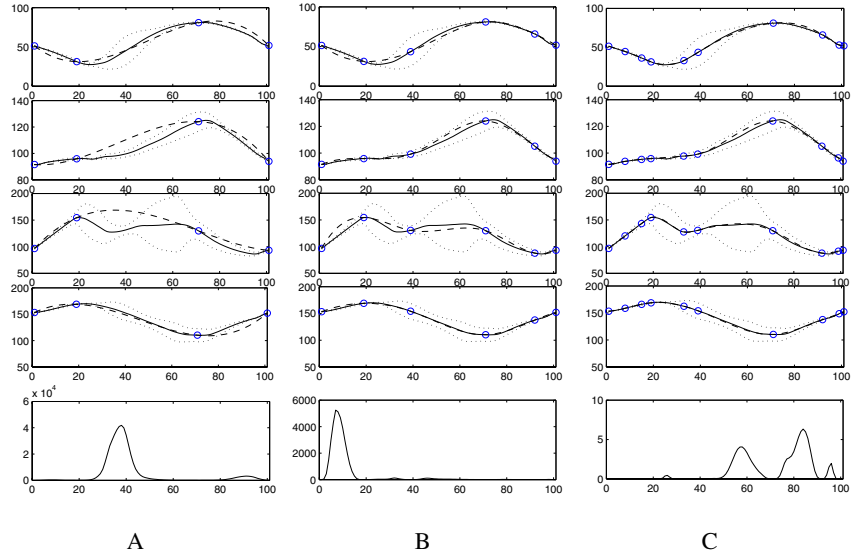
Fig. 5. Reconstruction and error. Solid: The mean gesture trajectory. Dashed: Interpolated cubic spline. Dotted: Variance of training data. Circles: Selected primitives and endpoints. A: With 2 primitives. B: With 4 primitives. C: With 8 primitives.
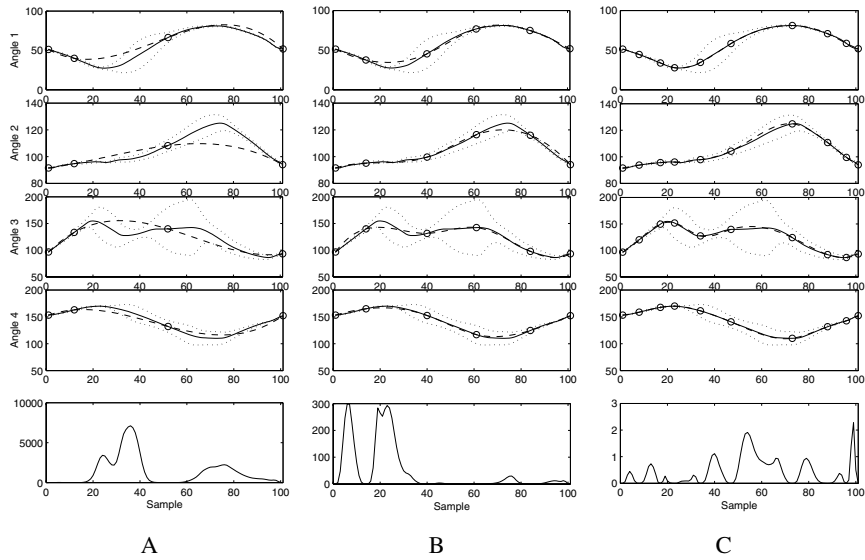
Fig. 6. Reconstruction and error (Optimizted version). Solid: The mean gesture trajectory. Dashed: Interpolated cubic spline. Dotted: Variance of training data. Circles: Selected primitives and endpoints. A: With 2 primitives. B: With 4 primitives. C: With 8 primitives.
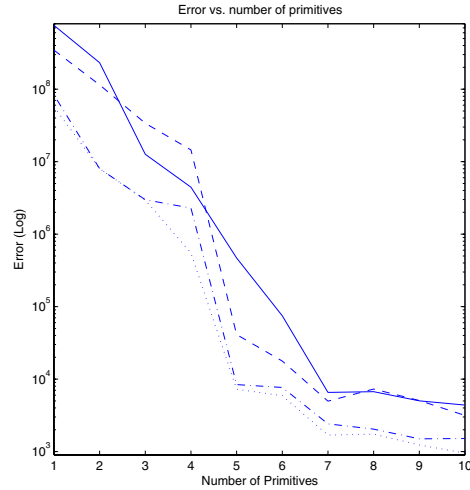
Fig. 7. Logaritmic scale of error vs. number of primitives. Solid: Reconstruction error after primitive selection without the density measure. Dashed: Reconstruction error after primitive selection with the density measure. Dash-dot: Reconstruction error after primitive selection without the density measure, but with optimization. dotted: Reconstruction error after primitive selection with the density measure and optimization.

each gesture, were collected after the reconstruction of the curves with the number of primitives ranging from 1-10. The error sums of both the optimized and none optimized version of our approach are plotted on a single logarithmic graph, shown in figure 7. The graph shows clearly that the optimized version has a lover error sum, but also that one or more of the reconstructions with four primitives were stranded in an unfortunate local minimum.

8 Conclusion

In this paper we have presented a framework for automatically finding primitives for human body gestures. A set of gestures is defined and each gesture is recorded a number of times using a commercial motion capture system. The gestures are represented using Euler angles and normalized. The normalization allows for calculation of the mean trajectory for each gesture along with the covariance of each point of the mean trajectories. For each gesture a number of primitives are found automatically. This is done by comparing the mean trajectories and cubic spline interpolated reconstructed trajectory by use of a error measurement based on density.

Our framework were implemented in two slightly different versions, were the slower proved to be superior, as it often is. Taken into consideration that our training data were very noisy, and the presented work is part of an ongoing research, we find the current results very promising, and will continue our work in this direction. We feel that the density measure have been proven as a factor that must be considered in

this line of work. Its is still hard to say exactly how many primitives are needed to get a natural reconstruction of a given gesture. But our tests indicate that somewhere between five and ten should be sufficient. It is obvious that other kind of curve-reconstruction techniques should result in much better reconstruction. But since the key-frames are to be used for recognition as well, it is important to have our key-frames at the points where the density is highest.

References

1. F. Bettinger and T.F. Cootes. A Model of Facial Behaviour. In *IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, May 17 - 19 2004.
2. A.F. Bobick. Movemnet, Activity, and Action: The Role of Knowledge in the Perception of Motion. In *Workshop on Knowledge-based Vision in Man and Machine*, London, England, Feb 1997.
3. A.F. Bobick and J. Davis. A Statebased Approach to the Representation and Recognition of GEstures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(12), 1997.
4. C. Bregler. Learning and Recognizing Human Dynamics in Video Sequences. In *Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997.
5. L. Campbell and A.F. Bobick. Recognition of Human Body Motion Using Phase Space Constraints. In *International Conference on Computer Vision*, Cambridge, Massachusetts, 1995.
6. R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley & Sons, Inc., 2 edition, 2001.
7. Hodgins et al. Segmenting Motion Capture Data into Dstinct Behaviors. *Unknown*, 2004.
8. J. Gonzalez. *Human Sequence Evaluation: The Key-Frame Approach*. PhD thesis, Universitat Autonoma de Barcelona, Barcelona, Spain, 2004.
9. N.R. Howe, M.E. Leventon, and W.T. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.
10. http://polhemus.com/. Polhemus, three-dimensional scanning, position/orientation tracking systems, eye tracking and head tracking systems., January 2005.
11. http://www.3dcrimescene.com/. Typical cold case reconstruction., January 2005.
12. O.C. Jenkins and M.J. Mataric. Deriving Action and Behavior Primitives from Human Motion Data. In *International Conference on Intelligent Robots and Systems*, Lausanne, Switzerland, Sep 2002.
13. A. Just and S. Marcel. HMM and IOHMM for the Recognition of Mono- and Bi-Manual 3D Hand Gestures. In *ICPR workshop on Visual Observation of Deictic Gestures (POINTING'04)*, August 2004.
14. A. Kale, N. Cuntoor, and R. Chellappa. A Framework for Activity-Specific Human Recognition. In *International Conference on Acoustics, Speech and Signal Processing*, Orlando, Florida, May 2002.
15. T.B. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3), 2001.
16. C. Rao, A. Yilmaz, and M. Shah. View-Invariant Representation and Recognition of Actions. *International Journal of Computer Vision*, 50(2), 2002.
17. C.R. Wren and A.P. Pentland. Understanding Purposeful Human Motion. In *International Workshop on Modeling People at ICCV'99*, Corfu, Greece, September 1999.