Human Motion Estimation from a Reduced Marker Set

Guodong Liu Jingdan Zhang Wei Wang Leonard McMillan
University of North Carolina at Chapel Hill*

Abstract

Motion capture data from human subjects exhibits considerable redundancy. In this paper, we propose novel methods for exploiting this redundancy. In particular, we set out to find a subset of motion-capture markers that are able to provide fast and high-quality predictions of the remaining markers. We then develop a model that uses this reduced marker set to predict the others. We demonstrate that this subset of original markers is sufficient to capture subtle variations in human motion.

We take a data-driven modeling approach to learn piecewise local linear models from a marker-based training set. We first divide motion sequences into segments of low dimensionality. We then retrieve a feature vector from each of the motion segments and use these feature vectors as modeling primitives to cluster the segments into a hierarchy of local linear models via a divisive clustering method. The selection of an appropriate linear model for reconstruction of a full-body pose is determined automatically via a classifier driven by a reduced marker set. After offline training, our method can quickly reconstruct full-body human motion using a reduced marker set without storing and searching the large database. We also demonstrate our method's ability to generalize over a variety of motions from multiple subjects.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-dimensional Graphics and Realism — Animation; I.2.10 [Vision and Scene Understanding]: Motion.

Keywords: motion capture, piecewise linear modeling, principal feature analysis, dimensionality reduction

1. Introduction

Motion capture, or mocap, is a prevalent technique for capturing and analyzing human articulations. Mocap has been widely used to animate computer graphics figures in motion pictures and in video games. However, most motion capture systems are cumbersome, expensive, intrusive, and time consuming. These drawbacks may not only prevent mocap data from being easy to use, but they also might make it impractical for other potential applications.

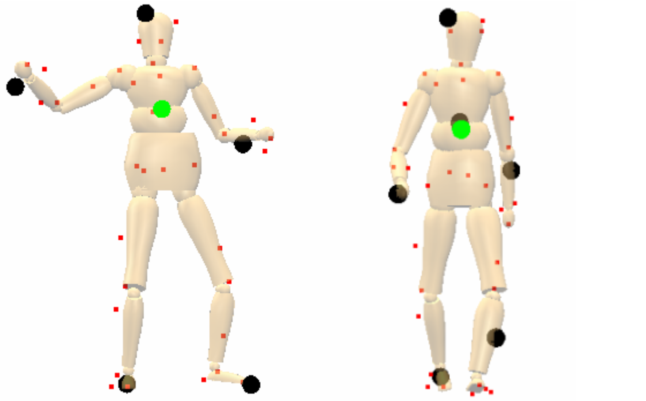


Figure 1: Shown above are the principal markers selected from 2 motion data sets. The principal markers are shown in black and the estimated markers are shown in red.

A common form of motion capture uses optical sensing of strategically placed *markers*. The subject often wears a black leotard in order to enhance the marker's contrast. A mocap system uses triangulation from multiple cameras to estimate the 3D position of each marker. Most often, the marker positions are converted to joint angles for an assumed skeletal model. Usually 40 to 50 markers are required to capture a motion sequence. More accurate motion recovery involves many more markers.

In this paper, we use a small set of markers, i.e., principle markers (Figure 1), but are still able to quickly generate plausible human motions on a frame-by-frame basis. This cheaper and faster motion capture system would benefit many applications, such as computer games and virtual reality environments where it is desired to have interactive, intuitive and accurate control of characters/avatars. In those applications, measurements from a few markers can be effectively used as control signals. Instead of wearing a tight Leotard with markers all over his/her body, a user may only need to wear normal cloth with only a few markers mounted on non-intrusive positions. Less mounting time also makes mocap feasible for more applications, since less overhead time is spent between users. Fewer marker measurements also reduce ambiguities during post-processing of mocap data, and thus require less human intervention.

There is considerable evidence that raw marker data significantly over-specifies the actual range of realistic human motions, and thus consistently exhibits redundancy and local linearity [Barbic et al. 2003, Grochow et al. 2004, Safonova et al. 2004, Chai and Hodgins 2005]. Motion data arising from similar motions can often be described by the same local linear model, which is valid for a limited range of articulation. We propose a data-driven approach to extract piecewise local linear models via divisive clustering from a motion database. We do not assume any explicit model (e.g., skeleton model). Instead, we use an implicit data-driven model based on dimensionality reduction and feature selection methods developed for data mining. Secondly, we

*email: {liug, zhangjd, weiwang, mcmillan}@cs.unc.edu

choose to carry out the modeling based on motion segments, instead of individual poses, i.e. frames. The reason is that a long and complex motion sequence can be considered as a concatenation from much simpler motion segments that often lie on a low dimensional linear space [Safonova et al. 2004; Grochow et al. 2004]. Motion segments provide a more appropriate resolution in terms of encapsulating the essence of human motion while retaining temporal relationship among neighboring poses. Finally, our model is very compact and completely eliminates the need for a motion database after offline training. It is also very fast in estimating motions from a reduced marker set. As the experiments show in later section, we can reconstruct human motion frame-by-frame at a rate of over 600 frames per second. Thus, our method shows great promise for use in most interactive motion applications.

The ultimate goal of our work is to provide fast and plausible estimates of a full-body pose based on a small marker set. We eventually hope to employ such methods for generating self-avatars for virtual environments (VEs), where the combined encumbrances of both head-mounted display and a full mocap setup are impractical. However, it is conceivable that a participant might undergo a short mocap session prior to entering a VE. This training data would then be used to estimate a plausible avatar from a significantly reduced marker set. An even better approach would allow a generic human-motion model to provide the pose estimates, perhaps adapting to the specific participant over time. In this paper we present our preliminary efforts to test the potential of such an approach.

2. Related Work

There has been extensive research on reusing motion capture data in animations, movies and interactive games. Most of them were on synthesizing motions by reordering the motion clips [Kovar et al. 2002; Arikan and Forsyth 2002; Lee et al. 2002; Pullen and Bregler 2002; Arikan et al. 2003], interpolating between motions [Rose et al. 1998; Kovar and Gleicher 2004, Mukai et al. 2005] and constructing models from the mocap data [Brand and Hertzmman 2000; Li et al. 2002]. On the other hand, Lee et al [2002] and recently Chai and Hodgins [2005] investigated the possibility of using low dimensional signals captured from human motions to control the characters/avatars. We also propose extracting a set of the most informative markers and their measurements, which can be used to drive an interactive application.

High-quality human motion data are very expensive to create and manipulate. A number of researchers have investigated the problem of inferring plausible human motion from a reduced measurement set. A common approach uses a small set of electromagnetic sensors to drive a virtual human model [Badler et al. 1993; Semwal et al. 1998; Kovar and Gleicher 2004; Grochow et al. 2004]. Recently Chai and Hodgins [2005] proposed a method for performance animation from a few marker measurements as control signals. Most of these methods typically assume prior knowledge of a skeleton model whose parameters, such as limb length, needs to be either accurately measured or estimated from the joint location [O'Brien et al. 2000]. They also rely on inverse kinematics to estimate the skeleton joint-angles from the constraints implied by the actual measurements. Various optimization and smoothing criteria are introduced to synthesize a full-body pose. Moreover, the selection of reduced measurement set is often determined by trial and error or intuition. We utilize Principal Feature Analysis (PFA) algorithm [Cohen et al. 2002] to systematically find a reduced marker set that maintains the most information while having less information overlapping among the selected markers. We also directly wok with marker data without assuming any skeleton model that is required for joint-angle-based data analysis. It is more natural to perform pose-to-pose distance calculations based on Euclidean distance metric for marker-based data than for the joint-angle-based data. For example, a small change in the angle of a shoulder may alter the shape of a pose much more than a similar change in a toe. As shown by Forbes and Fiume [2005], weighted joint-angle based representation of motion is more appropriate for measuring the similarity of poses. However, it is a nontrivial problem to design an appropriate weighting scheme for all joint angles. We apply linear regression instead of nonlinear optimization in reconstructing full-body poses from a reduced marker set. This results in faster reconstruction because the linear regressors are trained offline, and the initialization and convergence problems with nonlinear optimization may be avoided.

There is a significant redundancy in motion data due to the spatial and temporal coherence in human behaviors. The recent work of Safonova et al. [2004] and Grochow et al. [2004] demonstrates that specific simple motions can be accurately described via a low-dimensional parameterization based on dimensionality reduction techniques. Safonova et al. [2004] presented a method to synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. Grochow et al. [2004] proposed an inverse kinematics approach that applied a global nonlinear dimensionality reduction on human motion data using a Gaussian Process Latent Variable Model [Lawrence 2004]. Their approaches work well with a small homogenous data set. However, for a large motion dataset with various types of motions, a global modeling approach like theirs might be very slow and might not suit the dataset well. We propose a piecewise linear approach to model motions as a hierarchy of local linear models and is better suited to describe a large heterogeneous motion dataset. The local linear modeling approach for dimensionality reduction was considered by Bregler and Omohundro [1995], Hinton et al. [1995] and further developed in Chai and Hodgins [2005]. Chai and Hodgins [2005] utilized temporal coherence of the control signals to accelerate the nearest neighbor search for similar poses and dynamically construct a local linear model for the pose to be estimated. In contrast, we first segment motion sequence into segments of much simpler human behaviors that can be described by local linear models. We then identify those local linear models with a clustering method.

There have been studies on motion segmentation. Among them, Rose et al. [1998] segmented a motion sequence into simple motion strokes characterized by the abrupt change of velocities during transitions in order to retrieve example motions. Then new motions can be synthesized by interpolations between example motions with radial basis functions in the parameter space. Barbic et al. [2003] presented three methods to segment a motion sequence into segments of distinct behaviors. In contrast, we not only segment motion sequences into short and simple motion segments using the probabilistic principal component analysis (PCA) approach of Barbic et al. [2003], but also model motion segments with local linear models and cluster them by their similarities. In studying similarities among motion sequences, Kovar and Gleicher [2004] described a method for automatically search for logically similar motion in a motion database by finding ``close" motions and then uses them as intermediaries to find more distant motions. We construct a model hierarchy of motion segments, with similar motion segments at the same or the neighboring leaf nodes.

In recent years, there have been studies on markerless motion capture in computer vision community, although so far the proposed methods are still generally slower and less accurate than a marker-based approach. Chu et al [2002] proposed using nonlinear spherical shells to extract a skeleton model and estimating joint angles from volume data of motion sequences. They applied a global nonlinear dimensionality reduction technique, Isomap [Tenenbaum et al. 2000], for both the removal of pose-dependent nonlinearities and extractable skeleton curve features for a captured human volume. Their procedure models one motion sequence for a specific subject at a time. On the other hand, our method can model a heterogeneous motion data consisting of various human behaviors from different subjects at the same time. Given a new motion sequence, we can quickly use an appropriate model to fiducially estimate the full-body poses. Agarwal and Triggs [2004] proposed a global nonlinear regression method (relevance vector regression) to estimate 3D human pose from silhouettes. In comparison, we first construct a hierarchy of local linear models corresponding to low-dimensional linear spaces, and then apply simple linear regression within each of the local linear models. As the size and heterogeneity of motion data increases, the piecewise local linear modeling approach would be more feasible and efficient than the global modeling approach.

3. Proposed Method

Our goal is to estimate human motions from a small set of the most informative markers, i.e. principal markers. We give a brief overview of our method here (Figure 2), with more details explained later.

Principal marker selection. Principal component analysis (PCA) [Roweis 1997] is one of the most popular methods for dimensionality reduction of a feature set. However, the principle components in the lower dimensional space are latent variables. We want to choose a subset of the original features, i.e. principle markers that contain most of the essential information. We adapt a PCA-based Principle Feature Analysis method [Cohen et al. 2002] to select a principle marker set.

Piecewise linear modeling. We first apply the Probabilistic PCA (PPCA) [Tipping and Bishop, 1999] to divide a motion sequence into simple motion subsequences of distinct behaviors and local linearity. These subsequences are referred to as *motion segments*. We then characterize each motion segment by a feature vector and use them as modeling primitives to construct a hierarchy of local linear models via a divisive clustering method. Similar motion segments are partitioned into the same cluster leaf that corresponds to a particular local linear model. Poses in a leaf cluster are used to compute a linear mapping function from a set of principal markers to the rest markers.

Training classifier. In order to use the local linear models and the associated mapping functions to estimate full-body poses from a principle marker set, we need to identify the most appropriate model with the data only on principle markers. A classifier is trained for this task. We label the poses with a local linear model ID and train a *Random Forest* classifier [Breiman 2001]. Random Forest is a well justified and widely used classification and regression method in machine learning.

Motion reconstruction. Given a new motion sequence with only measurements from the principal markers, we associate each frame with the most appropriate local linear model using the *Random Forest* classifier and then use the associated mapping function to reconstruct full marker positions of the poses from the principal marker set. We smooth out possible discontinuities due to piecewise linear modeling by using a mixture of linear models for the poses at the transition between two linear models.

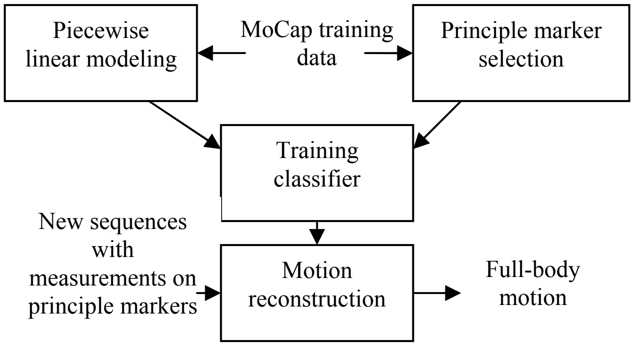

Piecewise linear modeling
MoCap training data
Principle marker selection
Training classifier
New sequences with measurements on principle markers
Motion reconstruction
Full-body motion

Figure 2: Human motion estimation process.

3.1 Principal Marker Selection

Throughout the paper, we treat each pose of motion data as a data point represented by a *3m*-dimensional column vector, $\mathbf{y} \in R^{3m}$, containing 3D marker positions of *m* markers. Thus a motion data set with *N* pose instances can be represented by a *3m*×*N* data matrix $\mathbf{Y}=[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$, where $\mathbf{y}_i$ is a column vectors of marker positions (*i=1*,…, *N*). For convenience, each of the 3D marker positions is referred to as a *feature* and then each pose can be considered as a high-dimensional data point with *3m* features.

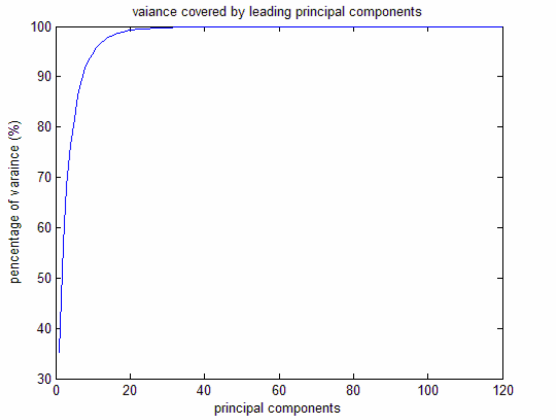


Figure 3: Percentage of variance explained by the principal components of a motion data set composed of 12670 frames with 40 markers (120 features). Fewer than 20 principal components are needed to reconstruct the original feature set to with 99% accuracy.

Motion data has significant redundancy, which can be demonstrated with PCA. Figure 3 shows the cumulative variance explained by the principal components for a data set comprised of a variety of human behaviors, including walking, running, bending, and washing. The first 10 principal components cover 95% of the variability, implying that a data set like this, with 40 markers (120 features), has only slightly more than 10 degrees of freedom. For selection of principle markers, an approach like PFA [Cohen et al. 2002] is very appealing, since it has comparable performance to PCA but selects a set of original features with a more intuitive interpretation than the principle components derived in PCA. PFA treats each feature as an individual measure, even though the three features of each marker are always measured together. So we design an algorithm based on PFA, but which selects a minimal set of principal markers instead of principal features in PFA.

3.1.1 Selection of principal markers

The basic idea of PFA is to exploit the structure of the principal components from PCA, and choose the principal features, which retain the essential information, in the sense of both maximizing variability of the features in the lower dimensional space and minimizing the reconstruction error. We first use PFA to partition all the features into clusters. Then we impose some criteria to weight the importance of each marker and select a minimal set of the most important markers satisfying a cover of all clusters of features. The steps of principal marker selection can be summarized as follows:

1. Run PCA on the covariance matrix of data **Y**.
2. Construct a $3m \times q$ matrix $\mathbf{A}_q$ by selecting the q dominant eigenvectors that are sufficient to satisfy a desired reconstruction error tolerance. The rows of $\mathbf{A}_q$ form the weight vectors, i.e. $\mathbf{V}_1, \ldots, \mathbf{V}_{3m}$, $\mathbf{V}_i \in R^q$ ($i=1:3m$), which are the projections of feature variables on the q leading principle components.
3. Take element-wise absolute value of $\mathbf{V}_i$ to obtain absolute weight vectors $|\mathbf{V}_i| \in R^q$ and use K-means clustering algorithm to partition the $3m$ absolute weight vectors to K clusters with K slightly greater than q.
4. Weight markers according to their importance. Remove the least important markers as long as every cluster is still covered by at least one marker after the removing.

A key rationale behind the feature clustering method is the realization that the rows of the matrix $\mathbf{A}_q$ can be used to effectively characterize the relationships between the features. In other words, if two features are highly correlated, they will have similar absolute value weight vectors.

We use the number of unique clusters containing a marker feature to define the importance of a marker. That is to say, a marker that appears in more distinct clusters is considered to be more important. To break ties between markers, we prefer those whose sum of the square distances from the marker's features to their cluster mean is minimal. Markers are sorted in the order of *least* importance. We continue removing the least important markers as long as every cluster is still covered. This process is repeated until no more markers can be eliminated.

3.1.2 Stability of Principal Marker Selection

K-means clustering is an iterative algorithm whose result depends on the choice of initial cluster means. However, it is our experience that the resulting clusters and the set of principal markers are surprisingly consistent and insensitive to the initial settings. In Figure 4 we illustrate a frequency histogram of the selected principal markers from 1000 runs of our principal marker selection algorithm on a motion data composed of 40 markers and 12,670 frames. All seven principle markers are consistently selected in more than 94% of all runs.

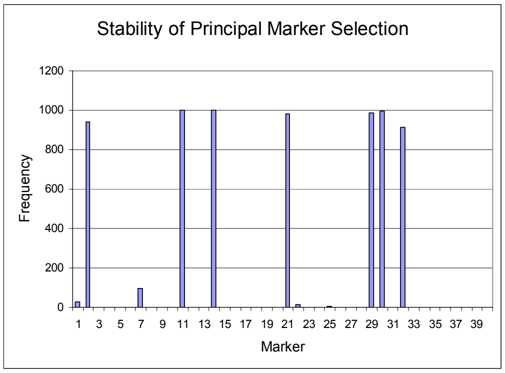


Figure 4: Frequency histogram of selected principle markers from 1000 runs.

3.2 Piecewise linear modeling

We propose a piecewise linear modeling approach to partitioning motion data into a collection of local linear models using motion segments as the modeling primitives. By keeping poses within one motion segment in one local linear model, we retain the temporal relationship among adjacent homogenous poses. Otherwise, poses from different behaviors could be clustered together, while temporally adjacent frames are clustered into different clusters, resulting in poor clustering and unnecessary model transitions that may cause jerkiness during the motion reconstruction phase.

3.2.1 Local linearity of human motions

Motion capture data exhibits a great deal of local linearity. Motion data of a specific behavior lies in a low-dimensional linear space. For example, motion poses of walking sequences comfortably fall into a two-dimensional linear space with over 90% of the variance covered by the two leading principal components. As Figure 5 shows, the dimension remains at two when we combine all the walking sequences, although the variance covered by the first two principal components dips a little bit due to stylish differences among sequences. This observation supports our hypothesis that human motion capture data lie in a piecewise local linear space, with motions of the same behavior or similar behavior (e.g., slow running and fast running) sharing the same local linear space.

3.2.2 Motion segmentation and characterization

Segmenting motions into simple and distinct behaviors may considerably facilitate the identification of local linear models. We choose the probabilistic PCA (PPCA) approach to segmentation [Barbic et al. 2004] since it was demonstrated to work fast and well with motion data and is easy to implement. PPCA treats motion data as an ordered sequence of poses (data points) and segments the motion where there is a local change in the distribution of the poses. In practice, a multivariate Gaussian distribution is assumed for a distinct behavior's poses, and PPCA is used to estimate their distribution. Motion segments from the same behaviors should have data rising from similar Gaussian distributions and sharing the same low dimensional space.

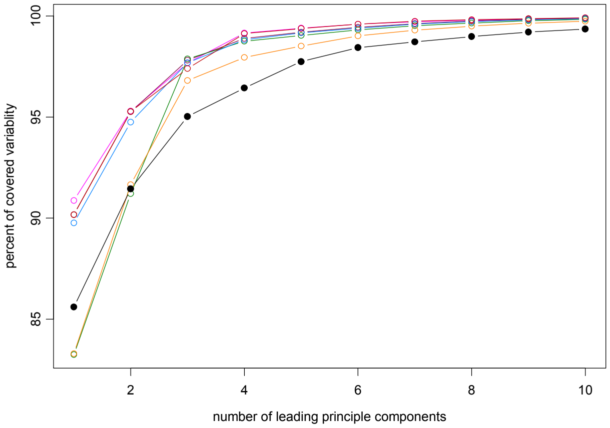


Figure 5: Percentages of variances explained by the principal components for different walking sequences. The curve in black with dots is for all the walking sequences combined, the rest curves with circles are for individual walking sequences of different styles.

Segmentation technique can divide complex sequences into simple distinct behaviors but provide no information on which segments are more similar to each other. We present here a divisive clustering method to identify and group segments that can be represented in the same low dimensional space, or in other words, by the same local linear model. Covariance matrix $\mathbf{\Sigma}$ and mean vector μ can uniquely determine a Gaussian distribution, we therefore form a feature vector, $\mathbf{f} = [w\mathbf{v}^T, (1-w)\mu^T]^T$, to characterize each of the motion segments, where $\mathbf{v}$ is a column vector consisting of the scaled variances and covariances. These values are retrieved by concatenating the elements in the upper triangle and the main diagonal of $\mathbf{\Sigma}$, and w is a weighting factor to balance the importance between the covariance matrix and the mean vector. If poses are data points in d-dimensional space, $\mathbf{\Sigma}$ is a $d\times d$ matrix and μ is a $d\times 1$ vector, then we have a $d(d+3)/2$-dimensional feature vector associated with the segment. This feature vector space has very high dimensionality with very sparse data points, with each segment considered as a data point, so we use PCA to reduce the dimensionality of the feature vectors. Our empirical results show that typically fewer than 40 principal components are needed to cover the 95% of the variance. Also, the difference in distribution between the segments is mainly due to the covariance matrix. Mean vector has very little impact.

3.2.3 *Local linear models and mapping functions*

We construct a hierarchy of local linear models by performing a divisive K-means clustering on feature vectors of segments with the Euclidean norm as a distance metric. If, at any level in the hierarchy, the distance between each feature vector and the cluster mean is within the established tolerance, the cluster is considered a leaf in the modeling hierarchy and a local-linear mapping function is computed for the cluster. Otherwise, the cluster is split into two child clusters, using a K-Means splitting algorithm with K set to 2. This splitting process continues until all clusters satisfy the distance tolerance.

After clustering is finished, each feature vector is uniquely partitioned into a leaf cluster, represented by a local linear model. Next, for each motion segment represented by a feature vector, all of its frames are labeled with the ID of the corresponding local linear model, i.e. cluster, where the feature vector is partitioned. For each local linear model, we compute a least squares mapping function to estimate measurements of the non-principal markers from a principal marker set. Assuming k out of m markers constitute a principal marker set, we represent a pose as a vector $\mathbf{y}=[\mathbf{x}^T, \mathbf{z}^T]^T$, where $\mathbf{y}\in R^{3m}$, $\mathbf{x}\in R^{3k}$ represents the 3D positions of principle markers, and $\mathbf{z}\in R^{3(m-k)}$ represents the 3D positions of the rest markers. Then the least squares mapping matrix $\mathbf{B}$ can be computed for a cluster of n poses as

$$\mathbf{B} = \mathbf{Z}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1},$$

where $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ is a $3k\times n$ matrix and $\mathbf{Z}=[\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n]$ is a $3(m-k)\times n$ matrix.

3.3 Random Forrest classifier

In order to use the local linear models and the associated mapping functions to estimate full-body poses from a principle marker set, a *Random Forest* [Breiman, 2001] classifier is trained to identify the most appropriate model with the data only on principle markers. In our classifier training process, for each frame labeled with its model ID, we use its principal marker values as input variables for Random Forest. *Random Forest* (RF) is a powerful classification tool that displayed outstanding performance in regard to classification error. RF grows and combines decision trees into predictive models. The overall prediction is determined by voting over all the trees in the forest and choosing the class with the most votes. Since the trees are generated randomly and independently, there is no risk of overfitting for large numbers of trees. As our experiment shows, RF performs well with a high degree of accuracy and is robust to the size and heterogeneity of the motion data. This, in turn, indicates that our piecewise linear modeling approach to label identification and selection method of principle markers is sufficient and effective to characterize motion data.

3.4 Motion reconstruction

3.4.1 *Estimation of poses*

Once we learn piecewise linear models and train Random Forest classifier with a training set, we are ready for estimation of poses from a principle marker set. Given measurements on principle marker set, denoted by vector $\mathbf{x}$, we use an RF classifier to identify the most appropriate local linear model and the associated least-squares mapping function. We then estimate the 3D positions of the remaining markers, $\mathbf{z}$, as $\mathbf{z} = \mathbf{B}\ \mathbf{x}$, where $\mathbf{B}$ is a mapping matrix.

3.4.2 *Estimating poses in Transition with Mixture of Local Linear Models*

An inherent shortcoming with piecewise linear modeling approach is the temporal discontinuity at the transitions between models, manifested as visible jerkiness in the reconstructed motion. We suspect that a change of bias in the reconstruction errors may be one of the leading causes to temporal discontinuity.

For example, if the reconstruction errors of consecutive frames are all biased towards the same direction, the motion may still appear smooth, although its root mean square (RMS) error may be a bit higher. On the other hand, if the biases are towards different directions, it may cause more severe jerkiness even if the RMS error is moderate. The bias tends to vary between models. Therefore, discontinuities (jerkiness) are frequently visible at the transitions between models.

We provide a simple metric to evaluate the jerkiness for each marker reconstruction. For a given marker, let $\mathbf{p}_t$ and $\mathbf{p}'_t$ be the true and predicted positions at time *t*; and $\mathbf{p}_{t-1}$ and $\mathbf{p}'_{t-1}$ be the positions at time *t-1*. Then we compute the true and predicted velocities $\mathbf{v}$ and $\mathbf{v}'$ as follows:

$$\mathbf{v} = \mathbf{p}_t - \mathbf{p}_{t-1}$$
$$\mathbf{v}' = \mathbf{p}'_t - \mathbf{p}'_{t-1.}$$

We then take the Euclidean norm of their vector difference (e.g. errors) as our jerkiness metric γ. i.e.

$$\gamma = \| \mathbf{v} - \mathbf{v}' \|.$$

When the errors are biased in similar directions, $\mathbf{v}$ and $\mathbf{v}'$ tend to be close to each other, leading to small values of γ**.** On the other hand, differences in the biases $\mathbf{v}$ and $\mathbf{v}'$ lead to larger values of γ.

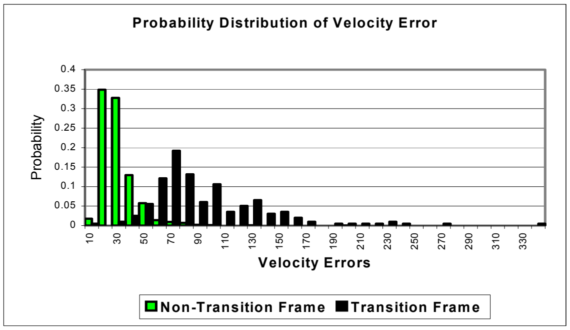


Figure 6: The probability distribution histogram of the velocity errors for reconstructed motions from a motion data set.

We verified the validity of this jerkiness metric experimentally on motion reconstruction using our method from a motion data set. In Figure 6, we plot a histogram of γ for all markers, where transition and non-transition frames are shown in different colors. The histogram shows that non-transition frames tend to be non-jerky, or smooth. However, nearly all jerky frames occur at transitions between clusters.

A typical solution is to incorporate a factor that evaluates the continuity of the pose relative to the previous poses into the optimization phase (Chai and Hodgins 2005; Grochow et al. 2004). We perform a fuzzy regression for the poses at the transitions of local linear models to smooth out the jerkiness. Instead of using only the local linear model where the pose is classified to reconstruct the full-body pose, we use a mixture of models associated with the current pose and some poses prior to it. This approach shares the spirit of fuzzy/soft classification and addresses the fact that transitional poses tend to be competed by different local linear models. Let $\mathbf{x}_t$ be a pose vector containing the 3D positions of principle markers at time *t*, we estimate the positions of the rest of the markers, $\mathbf{z}_t$, as

$$\mathbf{z}_t = \Sigma_i \, w_i \, \mathbf{B}_i \, \mathbf{x}_t,$$

where $w_i = r_i / (h+1)$ is a weight for the *i*th model, r_i is the number of poses classified to the *i*th model among the prior *h* poses and current pose, and $\mathbf{B}_i$ is the mapping matrix for the *i*th model. Basically, we want to put more weights on the model that is favored by more of the *h* poses prior to the current pose. In our experiments, *h*=10-30 works very well.

4. Experiments

4.1 Design

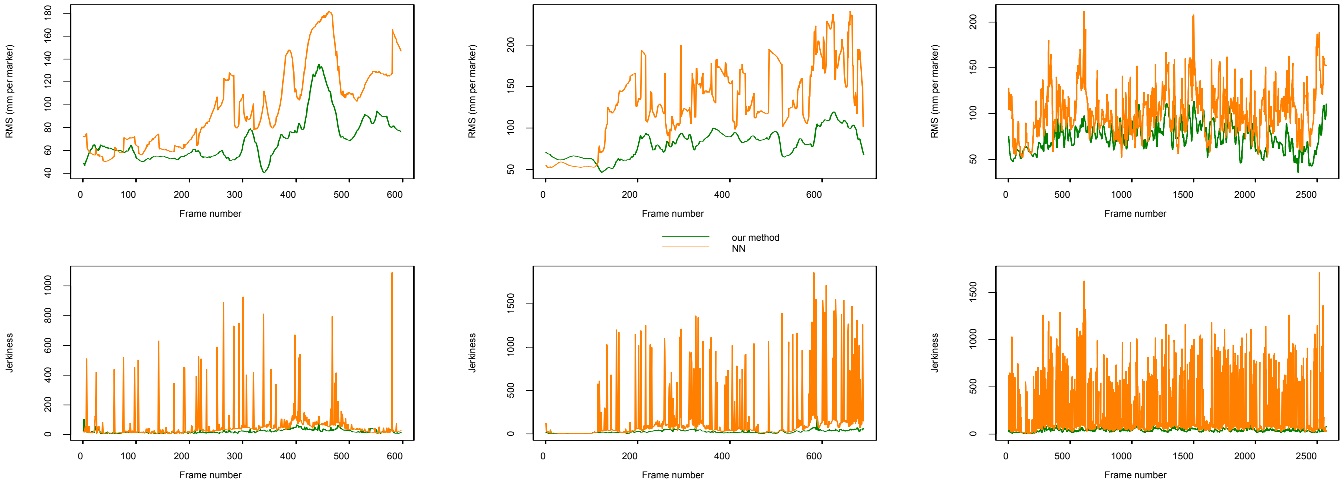
We evaluated our modeling approach using Carnegie Mellon University's Graphics Lab Motion capture database available at http://mocap.cs.cmu.edu. We used the motion data on a marker set with 41 markers. Typical motion data is captured in an absolute world coordinate frame. Our model, however, describes relative motion in a model-rooted frame. Therefore, a normalization step is required. To normalize data, we choose the marker located at the C7 vertebrae as the origin. The z-axis coincides with the z-axis of the original world coordinate frame. We compute a vector from the left shoulder marker to the right shoulder marker. We then project it to the horizontal plane and use the projected vector as the x-axis. The cross product between z and x axes produces the y-axis. There is no further normalization on the skeleton, such as normalization on the bone length.

To obtain a reasonable representation of motion data space, we prepared a large and heterogeneous human motion database including various motions from multiple subjects. We divided the motion sequences into a training set and a testing set, with the training set having similar sequences to the sequences in the testing set. We used the training set to learn piecewise linear models and extract a set of principle markers. Full-body poses were then reconstructed for the testing sequences based on the principle marker set and compared with the actual full marker measurements. We also estimated full-body poses by nearest neighbor search using the measurements from principal markers and compared the performance of our method to the nearest neighbor search.

4.2 Results

Our training set consists of 132 sequences with total 151,882 frames collected from 21 subjects. The training sequences contain a variety of motions, such as walking, running/jogging, golfing, soccer kicking, Salsa dancing, jumping, cartwheeling, climbing steps, etc. Even for the same category of motions, sequences of different styles from different subjects are included. The testing set contains 28 sequences with 19,553 frames from 18 subjects. Among them, there are 9 walking sequences, 6 running, 5 golfing, 2 cartwheeling, 2 Salsa dancing, 1 walking on uneven terrain, 1 running jump, 1 soccer kick, and 1 climbing three steps. Four testing sequences, namely, 1 walking, 1 soccer kicking, 1 running and 1 golfing are from 4 new subjects who never performed any motion that is used in the training set.

In selecting a set of principle markers from the training set, we computed PCA to cover 95% of the total variance. Then we used our principal marker selection method to select a set of six principal markers, placed at left forehead (LFHD), right elbow (RELB), left arm (LARM), right leg (RLEG), left toe (LTOE) and right toe (RTOE). The training set sequences were segmented into



(a) Soccer Kicking

(b) Running

(c) Salsa Dancing

Figure 7: Comparison of our method to Nearest Neighbors method in estimating three motion sequences. The top row compares the reconstruction RMS error (mm/marker). The second row compares the jerkiness (i.e., velocity error γ as previously defined).

271 segments with length varying from 128 to 3,670 frames (mean: 560; standard deviation: 425; median: 440). The dimensionality of motion segments could be as low as 2 for walking motion or as high as 14 for Salsa dancing. Hierarchical clustering of segments according to their feature vectors yielded 65 clusters, i.e., 65 local linear models. Principle marker positions were used to classify the frames into the local linear models via Random Forest. The classification error rate was 0.29%.

The reconstructed motions were visually plausible for all the testing sequences, with part of the results shown in the accompanying video. There was no visible jerkiness at the transitional poses. Our method performed reasonably well for the motions acted out by new subjects who were not in the database. We compared our method to nearest neighbors search method with respect to root mean squared (RMS) reconstruction errors and jerkiness. In general, our method created much more accurate results with less jerky estimation of motion. The average RMS error and velocity error over all the testing sequences are 45 mm/marker and 20 with our method, the corresponding errors with the nearest neighbor search are 56 mm/marker and 76 respectively. Figure 7 showed frame-by-frame RMS error and jerkiness for three testing sequences. The RMS error curve was much smoother using our method than nearest neighbor search. In other words, the nearest neighbor search method had a lot of spikes in the reconstruction error curve, which could indicate severe jerkiness, confirmed by its frame-by-frame jerkiness curve. In fact, our method reduced the jerkiness by 80% in most of the sequences. Visual inspection of the reconstruction results also confirmed our conclusion.

We also demonstrated in our experiments that the motion reconstruction was very fast. With the Random Forest classifier, the classification time was 0.00012 sec/frame; while the linear pose reconstruction time was 0.0014 sec/frame. This brings the total time needed for estimating a pose from a set of principal markers to 0.0015 sec/frame, or over 600 frames per second. We ran our experiments in Matlab V7 on a Dell Inspiron Laptop, with 1.4GHz CPU and 512M physical memory. A more powerful computer and more efficient code implementation may push the performance higher.

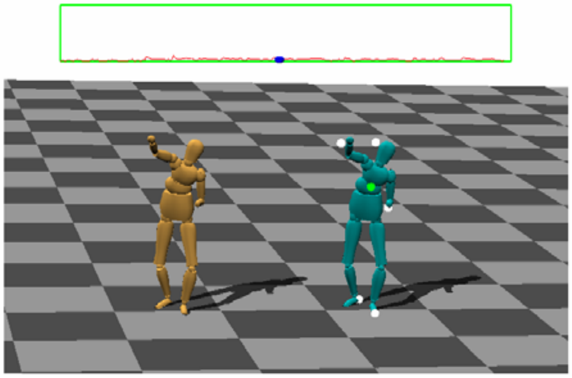


Figure 8: Shown above is a snapshot from our motion model viewer. The golden model on the left represents the actual pose data. The cyan model on the right shows an estimate of this pose based on the principal markers, which are depicted as white disks. The green disk indicates the origin marker. An RMS error meter for the entire marker set appears above the models with a full-scale value of 200 mm/marker. For more visualizations of our modeling results please view the accompanying video.

5. Discussion and Future Work

We presented a piecewise linear modeling approach to human motion data that are parameterized by a set of principal markers. We learned local linear models and principle markers from a training set of data samples. The motion reconstruction process is efficient with no need to search in a database. The experimental results demonstrated that our method can quickly generate plausible human motions on a frame-by-frame basis and scales well with size and heterogeneity of motions. Thus, we believe it is possible to use a few markers as control signals for interactive computer applications.

We identify a low-dimensional and local linear space at the motion segment level instead of the frame level. Motion segments offer a more appropriate resolution for motion data modeling. It retains temporal relationship to some extent by grouping temporally adjacent yet spatially homogenous frames together into one local linear model. Fewer local linear models are needed when modeling with motion segments than with frames, resulting in a more compact model hierarchy. It also improves the reconstruction quality by reducing unnecessary model transitions, a primary source of temporal discontinuity, i.e. jerkiness.

Our choice to model human motions in the marker space pushes motion data processing a step closer to raw data measurements, eliminating skeleton estimation, skeleton calibration and potential information loss during the conversion of marker measurements to joint angles. On the other hand, there may be a normalization issue with the use of marker data due to size differences among subjects. Nevertheless, our experiments showed that the performance of the proposed method was not sensitive to the normal variation in subjects' sizes. In the experiments, equivalent motions from different subjects tend to lie in the same local linear space, so the corresponding mapping function is actually computed based on data from different subjects. Calibration of subjects of different sizes does not appear to be essential with our marker-based approach. However, more experiments are needed in this regard.

We presented an algorithm for selection of a principle marker set. People may also want to follow their intuition or experience to select the principle markers, for example, on the extremities. It is of interest to compare the results obtained from automatically selected markers with those from the manually selected markers. Missing markers are often encountered in mocap data. It is desirable to use a training set with complete and precise marker measurements to learn a reliable model. However, in reconstruction of a new motion sequence, our method potentially allows for missing principle markers because Random Forest has efficient imputing method to replace missing values. It is worth conducting experiments to see to what extent the missing markers are allowed to retain an acceptable motion reconstruction.

Acknowledgments

The data used in this project was obtained from CMU's motion capture database (mocap.cs.cmu.edu), which was created with support from NSF EIA-0196217. We also would like to thank Mary Whitton for her voice over on the video.

References

AGARWAL, A., AND TRIGGS, B. 2004. 3D Human Pose from Silhouettes by Relevance Vector Regression. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR2004)*, 1063-1069.

ARIKAN, O., AND FORSYTH, D. A. 2002. Interactive motion generation from examples. In *Proc. SIGGRAPH 2002*, 483-490.

ARIKAN, O., FORSYTH, D. A., AND O'BRIEN, J. F. 2003. Motion synthesis from annotations. In *Proc. SIGGRAPH 2003*, 402-408.

BADLER, N., HOLLICK, M., AND GRANIERI, J. 1993. Real-time control of a virtual human using minimal sensors. *Presence*, 2(1):82-86.

BARBIC, J., SAFONOVA, A., PAN, J., FALOUTSOS, C., HODGINS, J., AND POLLARD, N. 2004. Segmenting Motion Capture Data into Distinct Behaviors. *Graphics Interface*, 185-194.

BREGLER, C., AND OMOHUNDRO, S. 1995. Nonlinear image interpolation using manifold learning. *In Advances in Neural Information Processing Systems*, 43-50.

BREIMAN, L. 2001. Random forest. *Machine Learning*, 45:5–32.

BRAND, M., AND HERTZMANN, A. 2000. Style machines. In *Proc. SIGGRAPH 2000*, 183-192.

CHAI, J., AND HODGINS, J. 2005. Performance animation from low-dimensional control signals. In *ACM Trans. Graph.* 24(3): 686-696.

CHU, C. W., JENKINS, O. C., AND MATARIC, M. J. 2003. Markerless kinematic model and motion capture from volume sequences, In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR 2003)*. 475-482.

COHEN, I., TIAN, Q., ZHOU, X. S., AND HUANG, T. S. 2002. Feature selection using principal feature analysis. *ICIP 2002*.

FORBES, K., AND FIUME, E. 2005. An efficient search algorithm for motion data using weighted PCA. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 67-76.

GROCHOW, K., MARTIN, S. L., HERTZMANN, A., AND POPOVIC, Z. 2004. Style-based inverse kinematics. In *Proc. SIGGRAPH 2004,* 522-531.

HINTON, G., REVOW, M., AND DAYAN, P. 1995. Recognizing handwriting digits using mixtures of linear models. In *advances in Neural Information Processing Systems*, 7:1015-1022.

KOVAR, L., GLEICHER, M., AND PIGHIN, F. 2002 Motion graphs. In *Proc. SIGGRAPH* 2002, 473-482.

KOVAR, L., AND GLEICHER, M. 2004. Automated extraction and parameterization of motions in large data sets. In *Proc. SIGGRAPH* 2004, 559-568.

LAWRENCE, N. D. 2004. Gaussian process latent variable models for visualization of high dimensional data. In *Advances in Neural Information Processing Systems,* 16:329-336.

LEE, J., CHAI, J., REITSMA, P., HODGINS, J., AND POLLARD, J. K. 2002. Interactive control of avatars animated with human motion data, In *Proc. SIGGRAPH 2002*, 491-500.

LI, Y., WANG, T., AND SHUM, H. Y. 2002. Motion texture: a two-level statistical model for character motion synthesis. In *ACM Transactions on Graphics,* 21(3):465--472.

MUKAI, T AND KURIYAMA, S. 2005 Geostatistical Motion Interpolation. *ACM Transactions on Graphics*, 24(3):1062-1070.

O'BRIEN, J. F., BODENHEIMER, B., BROSTOW, G., AND HODGINS, J. 2000. Automatic joint parameter estimation from magnetic motion capture data. In *Proceedings of Graphics Interface 2000*, 53-60.

PULLEN, K., AND BREGELER, C. 2002. Motion capture assisted animation: Texturing and synthesis. In *Proc. SIGGRAPH 2002,* 501-508.

ROSE, C., COHEN, M. F., AND BODENHEIMER, B. 1998. Verbs and adverbs: Multidimensional motion interpolation. In *IEEE Computer Graphics and Applications*. 18(5):32-40.

ROWEIS, S. 1997. EM algorithms for PCA and SPCA. *Advances in neural information processing systems 10*, 626–632. MIT Press.

SAFONOVA, A., HODGINS, J. AND POLLARD, N. P. 2004. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. In *Proc. SIGGRAPH 2004,* 514-521.

SEMWAL, S., HIGHTTOWER, S., HIGHTTOWER, R., AND STANSFIELD, S. 1998. Mapping algorithms for real-time control of an avatar using eight sensors*. Presence* 7(1):1-21.

TENENBAUM, J. B., DE SILVA, V., AND LANGFORD, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

TIPPING, M. E., AND BISHOP, C. M. 1999. Probabilistic Principal Component Analysis. *JRSS*B, 61(3):611–622.