# A Data-Driven Approach to Quantifying Natural Human Motion

Liu Ren[1] Alton Patrick[2] Alexei A. Efros[1] Jessica K. Hodgins[1] James M. Rehg [2]

[1]Carnegie Mellon University [2]Georgia Institute of Technology

Figure 1: Examples from our test set of motions. The left two images are natural (motion capture data). The two images to the right are unnatural (badly edited and incompletely cleaned motion). Joints that are marked in red-yellow were detected as having unnatural motion. Frames for these images were selected by the method presented in [Assa et al. 2005].

## Abstract

In this paper, we investigate whether it is possible to develop a measure that quantifies the naturalness of human motion (as defined by a large database). Such a measure might prove useful in verifying that a motion editing operation had not destroyed the naturalness of a motion capture clip or that a synthetic motion transition was within the space of those seen in natural human motion. We explore the performance of mixture of Gaussians (MoG), hidden Markov models (HMM), and switching linear dynamic systems (SLDS) on this problem. We use each of these statistical models alone and as part of an ensemble of smaller statistical models. We also implement a Naive Bayes (NB) model for a baseline comparison. We test these techniques on motion capture data held out from a database, keyframed motions, edited motions, motions with noise added, and synthetic motion transitions. We present the results as receiver operating characteristic (ROC) curves and compare the results to the judgments made by subjects in a user study.

**CR Categories:** I.3.7 [Computer Graphics]: Three Dimensional Graphics and Realism—animation I.2.6 [Artificial Intelligence]: Learning— Parameter learning G.3 [Mathematics of Computing]: Probability and Statistics—Time series analysis

**Keywords:** human animation, natural motion, machine learning, motion evaluation

## 1 Introduction

Motion capture is an increasingly popular approach for synthesizing human motion. Much of the focus of research in this area has been on techniques for adapting captured data to new situations. Motion capture data can be reordered in time [Arikan and Forsyth 2002; Kovar et al. 2002; Lee et al. 2002], similar motions can be interpolated [Wiley and Hahn 1997; Rose et al. 1998; Kovar and Gleicher 2004], motion can be edited [Gleicher 2001], and new motions can be generated by combining motions for individual limbs [Ikemoto and Forsyth 2004]. Models of human motion can also be used to synthesize new motion [Brand and Hertzmann 2000; Li et al. 2002]. Each of these techniques proposes heuristics or models that attempt to restrict the output of the algorithms to natural-looking motion, but no single naturalness measure exists to assess the quality of the output. In this paper, we explore whether it is possible to provide such a measure.

How can we quantify what it means for a sequence of human motion to appear natural? One approach is to propose a set of heuristic rules that govern the movement of various joints. If a given sequence violates any of the rules, it is judged to be unnatural. For example, a character's motion could be tested for angular momentum conservation in flight or violation of the friction cone when the foot is in contact. This bottom-up approach will likely have difficulty with the more stylistic elements of human motion, because a motion can be physically correct without appearing natural.

A second approach is to develop a set of perceptual metrics that provide guidelines for the flaws that people are likely to notice [Reitsma and Pollard 2003; O'Sullivan et al. 2003; Harrison et al. 2004; Pollick et al. 2003]. For example, Reitsma and Pollard measured the sensitivity of users to changes in horizontal and vertical velocity. Taken together, such studies could provide a set of guidelines to assess whether a given motion will be perceived as natural.

A third approach is to train a classifier to distinguish between natural and unnatural movement based on human-labeled, ground-truth data [Ikemoto and Forsyth 2004; Wang and Bodenheimer 2003]. For example, Wang and Bodenheimer used an optimization approach to find weights for a transition metric that best matched the judgments of sequences by users. Here we present an alternative take on this approach. We assume that the learning algorithm will be trained only on positive (natural) examples of the motion. We make this assumption because natural motions are readily available from commercial motion capture systems. Negative (unnatural) examples, on the other hand, are precious because each must be hand labeled by a person. As a consequence of this scarcity, the negative

[1]{liuren|efros|jkh}@cs.cmu.edu
[2]{apatrick|rehg}@cc.gatech.edu

examples that would be required for training do not exist. A further concern is that the characteristics of these negative examples are likely to be specific to the adaptation method that generated them and not representative of unnatural motions in general. We will demonstrate that using our approach, a variety of motions can be assessed using models that have been trained on a large corpus of positive examples (Figure 1).

Our approach to this problem is based on the assumption that the evaluation of naturalness is not intrinsically a subjective criterion imposed by the human observer but is, instead, an objective measure imposed by the data as a whole. Simply put, motions that we have seen repeatedly are judged natural, whereas motions that happen very rarely are not. Humans are good at this type of evaluation because they have *seen a lot of data*. The amount of collected motion capture data has grown rapidly over the past few years and we believe that there is now an opportunity for a computer to analyze a lot of data, resulting in a successful method for evaluating naturalness.

The contributions of this paper are threefold. First, we pose the question of whether it is possible to quantify natural human motion independent of any specific adaptation task. Second, we hierarchically decompose human motion into its constituent parts (individual joints, limbs, and full body) and build a statistical model of each one using existing machine learning techniques. We then combine these models into an ensemble model for classification of the motion as natural or unnatural. We present ROC curves of the performance of these techniques on a broad set of test sequences and compare the results to human performance in a user study. And finally, we contribute a substantial database of human motion and a testing set that will enable others to apply their algorithms to this problem. Both training and testing datasets are freely available on the web: http://graphics.cs.cmu.edu/projects/natural/.

## 2 Related Work

To our knowledge there is little work in computer animation that directly explores the question of quantifying natural human motion; however, many algorithms for synthesizing and editing human motion have been designed with the goal of restricting their output to natural human motion. We briefly review that work and then discuss related problems in other disciplines.

One early technique for amplifying the skills of the naive animator was Perlin's work using modulated sine waves and stochastic noise to create lifelike animation [Perlin 1995]. We test on both positive and negative sequences that are similar in that sinusoidal noise has been added to motion capture data.

Many motion editing techniques have been proposed, each with a set of optimization criteria intended to ensure that the resulting motion is natural (see, for example [Gleicher 2001; Sulejmanpasic and Popovic 2005]). Some of these techniques have been adapted into commercial software, and we use Maya to perform editing on motion capture data to generate part of our negative test set.

Motion graphs create new animations by resequencing pieces of motion capture data. The naturalness of the resulting motion depends largely on the quality of the motion transitions. Several algorithms have been proposed for creating natural transitions [Lee et al. 2002; Kovar et al. 2002; Arikan and Forsyth 2002]. We use synthetic motion transitions, both good and bad, as part of the test set in our experiments.

Wang and Bodenheimer [2003] used optimization to tune the weights of a transition metric based on example transitions classified by a human viewer as good or bad. They made several assumptions to make the optimization process tractable. For example, they did not consider how changes in the blending algorithm would affect the naturalness for a given distance metric. They also studied the optimal duration for a transition given a previously learned distance measure [Wang and Bodenheimer 2004].

Limb transplant is one way to generalize the motion in an available database. Ikemoto and Forsyth [2004] used an SVM to classify a synthesized motion as "looks human" or "does not look human." Their approach was quite effective for this problem, but it is a supervised learning approach and therefore requires a relatively large number of positive and negative training examples specific to limb transplant. In contrast, our goal is to use unsupervised learning to construct a measure that can be trained only on positive examples and that works for motion produced by a variety of motion editing algorithms.

The question of how to quantify human motion is also related to research that has been performed in a number of other fields. For example, researchers interested in speaker identification have looked at the problem of deciding whether a particular speaker produced a segment based on a corpus of data for that speaker and for others [Cole 1996]. Classifying natural vs. unnatural images for fraud detection is similarly related to our problem [Farid and Lyu 2003].

Closer to our problem is the work of Troje [2002] who was interested in identifying features of a human walk that can be used to label it as male or female. He reduced the dimensionality of the dataset as we do, with PCA, and then fit sinusoids to the resulting components. This approach is specific to a cyclic motion such as walking and would not easily generalize to our very large, heterogeneous database. However, the performance of his classifier was better than that of human subjects on a point light visualization of the walking motion.

Researchers working in activity recognition have looked at detection of unusual activities, which is similar to our problem in that an adequate negative training set would be difficult to collect. As a result, most approaches have focused on unsupervised learning. For example, Zhong and his colleagues [2004] used an unsupervised learning approach to detect unusual activity in video streams of human motion. Hara and his colleagues [2002] took motion detector data acquired from an intelligent house, performed vector quantization, and estimated the probability of a sequence of sensor data with a HMM. Hamid and his colleagues [2005] used clustering of event n-grams to identify and explain anomalous activities.

## 3 Data

We explore the performance of three classes of statistical machine learning techniques when trained on a large database of motion capture data and tested on sequences of unnatural and natural motion from a number of different sources. Because the validity of these results depends heavily on the training and testing datasets, we first describe those datasets and then explain the statistical techniques and show their performance.

### 3.1 Training Database

The training database consisted of 1289 trials (422,413 frames or about 4 hours) and included motions from 34 different subjects performing a variety of behaviors. Those behaviors included locomotion (42%: 5% jumping, 3% running, and 33% walking), physical activities (16%: basketball, boxing, dance, exercise, golf, martial

arts), interacting with the environment (7%: rough terrain, playground equipment), two subjects interacting (6%), and common scenarios (29%: cleaning, waiting, gestures).

The motion was originally captured with a Vicon motion capture system of 12 MX-40 cameras [Vicon Motion Systems 2005]. The motion was captured at 120Hz and then downsampled to 30Hz. The subjects wore 41 markers, the 3D positions of which were located by the cameras. Using an automatically obtained skeleton for the user, the motion was further processed to the ASF/AMC format, which includes absolute root position and orientation, and the relative joint angles of 18 joints. These joints are the head, thorax, upper neck, lower neck, upper back, lower back, and left and right humerus, radius, wrist, femur, tibia, and metatarsal.

For the experiments reported here, we converted each frame of raw motion data to a high-dimensional feature vector of angles and velocities. For the root segment, we compute the angular velocity and the linear velocity (in the root coordinate system of each frame). For each joint, we compute the angular velocity. The velocities are computed as a central difference between the joint angle or the position on the previous frame and on the next frame. As a result, both joint angles and their velocities can be represented by unit quaternions (four components each). The complete set of joint angles and velocities, together with the root's linear velocity (three components) and angular velocity (quaternion, four components), form a 151-dimensional feature vector for each frame. The quaternions are transformed to be on one-half of the 4D sphere to handle the duplicate representation of quaternions. If the orientation of a joint crosses to the other half-sphere, we choose the alternative representation for that quaternion and divide the motion sequence at the boundary to create two continuous sequences. Fortunately this problem occurs relatively rarely in natural human motion because of human joint limits.

### 3.2 Testing Motions

We generated a number of different test sets in an effort to span the space of natural and unnatural motions that might be generated by algorithms for producing human motion. Unlike our training data, the testing suite contains both positive (natural) and negative (unnatural) examples.

The negative testing sequences were obtained from a number of sources:

- Edited motions. Alias/Wavefront's Maya animation system was used to edit motion capture sequences to produce negative training examples. The editing was performed on either a joint or a limb using inverse kinematics.
- Keyframed motions. These motions were keyframed by an animator with significant Maya experience but limited keyframing experience.
- Noise. Noise has been used to generate human motion [Perlin 1995] and to improve the quality of captured motion by adding variation. We generate both positive and negative testing examples by varying the amount of noise and relying on a human observer to assess the naturalness of the motion.
- Motion transitions. These motions were computed using a commonly accepted metric for transitions (maintain contact constraints and keep the sum of the squared changes in joint angles below a threshold). Transitions above a high threshold and below a low threshold were then classified as good or bad by a human viewer.
- Insufficiently cleaned motion capture data. In the process of cleaning, motion capture data is transformed from the 3D marker locations to relative joint angles using a model of the subject's skeleton. For most marker sets, this process is accomplished through the use of inverse kinematics. If the markers have not been placed carefully or the kinematic chain is near a singularity, this process may result in unnatural motion (for example, knees that do not fully extend or swing out to the side if significantly bent).

The negative, or unnatural, testing set consisted of 170 trials (27774 frames or 15 minutes).

The positive tests consisted primarily of motion capture data that was held out from the database. Additional positive testing data were created by adding noise to these motions and by generating motion transitions that were judged good by an expert human viewer. The natural motions consisted of 261 trials (92377 frames or 51 minutes).

## 4 Approach

The input data for our models, motion capture data, is a multivariate time series consisting of vectors of features (joint angles and velocities) sampled at discrete time instants. From this perspective, a model for natural motion must capture probabilistic dependencies between features across time. We construct this model in three steps. First, we select a statistical model to describe the variation in the data over time. We investigate three relatively standard techniques: mixtures of Gaussians (MoG), hidden Markov models (HMM) and switching linear dynamic systems (SLDS). Associated with each model is a set of model parameters and a likelihood function that measures the probability that an input motion sequence could be generated by the model. Second, we fit the model parameters using a corpus of natural human motion as training data. Third, given a novel input motion sequence, we compute a score which can be interpreted as a measure of naturalness.

By thresholding the naturalness score we obtain a classifier for natural motion. There are two types of classification errors: false positives (the classifier predicts natural when the motion is unnatural) and false negatives (the opposite case). By varying the threshold we can trade-off these two types of errors. The *ROC curve* for a classifier summarizes its performance as a function of the threshold setting [Van Trees 1968] (see Figures 3 and 4 for examples). Each threshold choice corresponds to an operating point on the ROC curve. By comparing the area under the ROC curve, we can measure the relative performance of a set of classifiers without the need to choose a particular threshold. In practice the choice of operating point on the ROC curve will be dictated by the application requirements and will be assessed using a set of positive and negative examples that were not used for training.

We could construct a single statistical model of naturalness using the full 151-dimensional input feature vector from Section 3.1 for training. However, learning an accurate model for such a high-dimensional feature vector is difficult, even with a (relatively) large amount of training data. Therefore, we propose to hierarchically decompose the full body motion into its constituent parts and train an *ensemble* of statistical models, each responsible for modeling a particular part: joints, limbs, or the whole body. Given an input sequence, these smaller models would produce a set of likelihood scores and an ensemble rule would be used to combine these scores into a single naturalness measure. The ensemble approach has three potential advantages over creating a single model based on the complete feature vector:

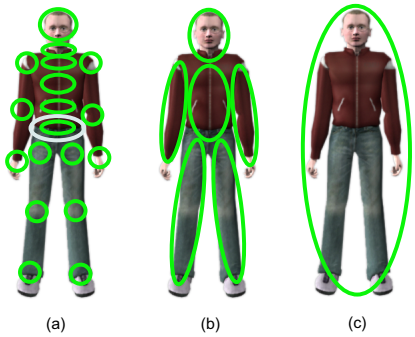


Figure 2: The three hierarchical groups of features. (a) At the lowest level each joint and its velocity form a feature group. Each feature group is illustrated as a green circle. The white circle represents the group of features from the root segment (linear velocity and angular velocity). (b) The next level consists of sets of joints grouped as limbs. (c) At the highest level, all the joints are combined into one feature group (without velocity information).

- One potential problem in learning the parameters of statistical models is overfitting, which occurs when a model has excessive capacity relative to the amount of available training data. When overfitting occurs, the trained models do not generalize well during testing because they are excessively tuned to the training data set. The ensemble approach gives us flexibility in controlling the capacity of the individual models to prevent overfitting. In particular it allows us to control the degree of coupling between features in the model.

- In some motion sequences, the patterns of unnatural motion may be confined to a small set of joint angles. These cases can be difficult to detect with a single statistical model, because the small set of features with unnatural motion will be swamped by the majority of the features which are exhibiting natural motion. The ensemble approach avoids this problem because our method of combining the statistical models looks for an unnatural classification by *any* of the models, not an average classification of unnaturalness.

- The ensemble approach makes it possible to examine small groups of joints and identify the ones most strongly associated with the unnatural motion. This property should make it possible to provide guidance to the animator about what elements of the motion deserve the most attention.

We designed groups of features to capture dependencies between joints at different scales. Each group of features forms a feature vector that is associated with a single model in the ensemble. Specifically, given the input 151-dimensional feature vector described in Section 3.1, we define a set of 26 smaller feature vectors by combining joint angles and joint velocities into groups of features (figure 2). At the lowest level, we create an 8-D feature vector from each of the 18 basic joints (angle and velocity). Another feature vector is created for the linear and the angular velocity of the root segment (seven features). To represent the aggregate motion of parts of the body, we assign a feature vector to each of the limbs: two arms (each three joints; 24 features), two legs (each three joints; 24 features), the head-neck group (head, upper neck, lower neck; 24 features) and the torso/root group (thorax, upper back, lower back, plus root; 31 features). Finally, at the top level, we define a feature vector representing the full body pose (rotation angles for all 18 joints but no velocities; 72 features). For the models created using HMM and SLDS, the feature vectors that comprise each of these feature groups are first processed with PCA (99% variance kept for the full-body model, 99.9% variance kept for the smaller models) to reduce the dimensionality.

Given an ensemble of models, we generate a naturalness measure for a motion sequence $D$ of length $T$ by first computing a score $s_i$ for each model, where the model has parameters $\theta_i$:

$$s_i = \frac{\log P(D \mid \theta_i)}{T}$$

The scores for each model will generally not be in the same range. Therefore we must normalize the scores before they can be combined. For each model, we compute the mean $\mu_i$ and standard deviation $\sigma_i$ of the scores for the training data (after eliminating a small percentage of the high and low scores to reduce the effect of outliers). The final score for sequence $D$ is then computed as follows:

$$s = \min_i \left( \frac{s_i - \mu_i}{\sigma_i} \right), i = 1, 2, \ldots, 26$$

We choose the minimum (worst) normalized score from among the $s_i$ because we assume that the entire motion should be labeled as unnatural if any of its constituent feature groups have bad scores.

We now describe the three statistical models used in our experiments, as well as a baseline method and a user study used for validating our results.

### 4.1 Mixture of Gaussians

We first experimented with a mixture of Gaussians (MoG) model because of its simplicity. The probability density of each feature vector was estimated using a mixture of 500 Gaussians, each with a spherical covariance. In this rudimentary representation, the dynamics of human motion are only encoded through the velocity components of the feature vector. As the result, this model is quite weak at modeling the dynamics of human movement.

### 4.2 Hidden Markov Models

Next, we experimented with a hidden Markov model (HMM) [Rabiner and Juang 1993], because it explicitly encodes dynamics (change over time) and has been shown to work extremely well in other time-series domains such as speech recognition. In a HMM, the distribution of the body poses (and velocities) is represented with a mixture of Gaussians. In general, each hidden state in a HMM indexes a particular mixture density, and transitions between hidden states encode the dynamics of the data. Given positive training examples, the parameters of the HMM can be learned using the Expectation-Maximization (EM) algorithm. The parameters consist of the probabilities in a state transition matrix for the hidden state, an initial state distribution, and mixture density parameters. In the general case, this set of parameters includes mixture weights for each hidden state and the mean vectors and covariance matrices of the Gaussians.

For the full body HMM, we used a model with 180 hidden states. For the other feature groups comprising the ensemble of HMM, we used only 60 hidden states because the feature vectors were much smaller. Each hidden state in the HMM was modeled as a single Gaussian with a diagonal covariance matrix.

### 4.3 Switching Linear Dynamic Systems

A switching linear dynamic system (SLDS) model can be viewed as a generalization of a HMM in which each switching state is associated with a linear dynamic system (LDS) instead of a Gaussian distribution over the output space [Pavlović et al. 2000]. In a HMM, each switching state defines a "region" in the output space (e.g, poses and velocities), where the mean vector determines the location of the region and the covariance matrix determines its extent. In contrast, each LDS component in an SLDS model defines a family of trajectories with linear dynamics. We used a second-order auto-regressive (AR) model in our experiments. In this model, trajectories begin at an initial state that is described by a mixture of Gaussians. As the trajectory evolves, the state of the motion at time $t$ is described by a linear combination of the state values at times $t-1$ and $t-2$ and the addition of Gaussian noise. By switching between these LDS components, the SLDS can model a system with nonlinear, non-Gaussian dynamics using a set of simple building blocks. Note that our application of SLDS does not require a separate measurement process, because we model the motion directly in the feature space.

Closely related to our SLDS model is the motion texture model [Li et al. 2002]. The primary difference is that the motion texture approach confines each LDS element to a "texton" that is constrained to begin and end at specific keyframes, whereas we adopt the classical SLDS framework where transitions between LDS models can occur at each time step.

As in the HMM case, the SLDS model parameters are estimated using the EM algorithm. However, a key difference is that exact inference in hybrid dynamic models like SLDS is generally intractable [Lerner 2002]. We employed an approximate Viterbi inference algorithm which computes an approximation to the highest probability switching sequence [Pavlović et al. 2000].

Given a new motion sequence, we compute a score that corresponds to the log likelihood of the data under the SLDS model. This score is the sum of the log likelihoods for each frame of data. Per-frame scores depend on the cost of switching between models and the size of the one-step-ahead error between the model's prediction and the actual feature vector.

For the full body SLDS, we used an SLDS model with 50 switching states. For the other groups of features comprising the ensemble model, we used 5 switching states each. We used diagonal covariance matrices for the noise process.

### 4.4 Naive Bayes (Baseline Method)

To establish a baseline for the other experiments, we also implemented a simple marginal histogram probability density estimator based on the Naive Bayes (NB) model. Assuming independence between the components of our 151-dimensional feature vector (which is clearly wrong), we computed 1D marginal histograms for each feature over the entire training database. Each histogram had 300 buckets. Given this model, we estimated the score of a new testing sequence by summing over the log likelihoods of each of the 151 features for each frame and then normalizing the sum by the length of the motion sequence. Note that this method captures neither the dependencies between different features (even those comprising a single joint angle), nor the temporal dependencies between features at different frames (although velocities do provide some measure of dynamics). As expected, this method does not perform particularly well, but we included it as a baseline with which to compare the other, more complicated approaches.

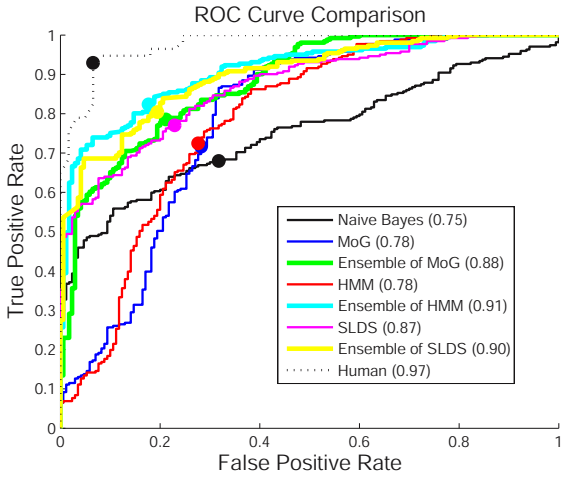


Figure 3: The ROC curves for each statistical model and for the human subjects in our user study. The circle on each curve represents the equal error rate. The area under the ROC curve is given in parentheses.

### 4.5 User Study

To evaluate our results, we performed a user study approved by Institutional Review Board (IRB) of Carnegie Mellon University. Twenty-nine male subjects and twenty-five female subjects with different backgrounds and races were obtained by university-wide advertising.

We randomly selected and rendered 118 motion sequences from our testing set (approximately half from the positive testing set and half from the negative testing set). We showed the rendered videos to subjects in two segments with a 10 minute break between the segments. Each segment contained half the sequences in a random order and the ordering of the presentation of the two segments was randomized between subjects. After watching each motion, the subjects wrote their judgment about the naturalness of the motion (yes or no). The total length of the study (including the break) was about 30 minutes. For comparison with the statistical models, the results of the user study are summarized in Section 5.

## 5 Experiments

We trained the statistical models on the database of four hours of human motion and tested them on a set of 261 natural and 170 unnatural motions. Figure 3 shows the ROC curves for each method. The ROC curve for the user study was computed by varying the threshold for the number of subjects who must mark a motion as natural for it to be labeled as natural. The testing set for the human subjects was only 118 of the 431 testing motions in order to prevent fatigue.

Table 1 gives the area under the ROC curve for each method. For the single full-body models (151 features), SLDS had the best performance, followed by HMM and MoG. Each ensemble of 26 models performed better than the single model that used the same statistical technique. This improvement occurs largely because the smaller statistical models and our method of combining their scores makes the ensemble more sensitive to unnatural motion of a single joint than a single statistical model. The ensemble of HMM had the largest area under the ROC curve, although the performance of

| Method | Positive Test Set (261) | Bad Motion Capture (37) | Edited (60) | Keyframed (11) | Noise (30) | Transition (32) | Area Under ROC | Number of Parameters |
|---|---|---|---|---|---|---|---|---|
| Naive Bayes | 0.69 | 0.75 | 0.73 | 0.80 | 0.76 | 0.40 | 0.75 | 45,600 |
| MoG | 0.71 | 0.86 | 0.97 | 1.00 | 0.37 | 0.28 | 0.78 | 76,000 |
| Ensemble MoG | 0.74 | 0.89 | 0.80 | 1.00 | 0.80 | 0.40 | 0.88 | 201,000 |
| HMM | 0.72 | 0.78 | 1.00 | 1.00 | 0.53 | 0.22 | 0.78 | 21,087 |
| Ensemble HMM | 0.82 | 0.89 | 0.78 | 1.00 | 0.83 | 0.75 | 0.91 | 43,272 |
| SLDS | 0.76 | 0.78 | 0.75 | 1.00 | 0.43 | 1.00 | 0.87 | 333,150 |
| Ensemble SLDS | 0.82 | 0.76 | 0.82 | 1.00 | 0.67 | 0.97 | 0.90 | 159,340 |
| Human Subjects | 0.93 | 0.75 | 1.00 | 0.81 | 1.00 | 0.92 | 0.97 | NA |

Table 1: The percentage of each type of testing data that was classified correctly by each classification method (using the point on the ROC curve with equal error rate). The number of test sequences for each type of motion is given in parentheses.

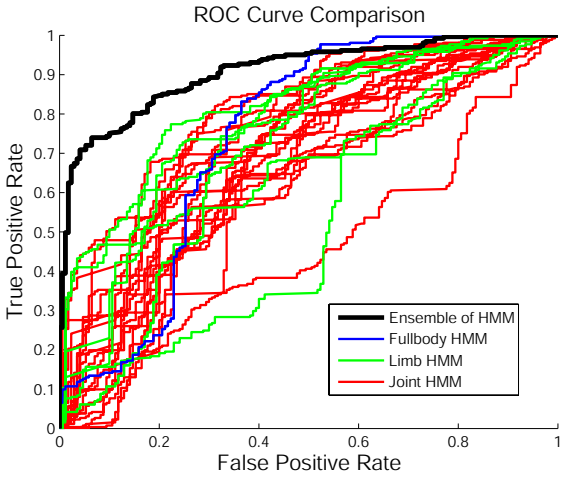


Figure 4: ROC curves for each of the 26 HMM and the combined ensemble HMM. The HMM for the individual joints are shown in red, for limbs in green, and for the full body in blue. The lowest curve corresponds to the right wrist which also causes the curve for the right arm to be low.

all three ensemble methods was similar. The human subjects performed significantly better than any of the methods, indicating that it may well be possible to develop better methods.

Table 1 also gives the percentage of the testing data that were classified correctly for each category of the test set and each model. The threshold setting for each classifier corresponds to the point of equal error rate on the ROC curve (see Figure 3). This point on the ROC curve is where the percentage of false positives equals the percentage of false negatives. Bad motion capture data was not easy for most of the classifiers to detect with only the ensemble of MoG and of HMM having success rates near 90%. The human subjects were also not particularly good at detecting those errors, perhaps because the errors were generally of short duration and the subjects did not have experience with the process of capturing or cleaning motion capture data. All of the methods were able to correctly classify more than 70% of the edited motions as unnatural, and MoG, HMM, and the human subjects had a success rate of over 95% on those motions. The keyframed motions were small in number and were largely classified correctly as unnatural by all methods and the human subjects. The addition of sinusoidal noise was more difficult for most of the methods to detect with only ensemble MoG and ensemble HMM achieving scores near 80%. The human subjects, on the other hand, could easily discriminate these motions, scoring 100%. Motions with bad transitions were the most difficult type

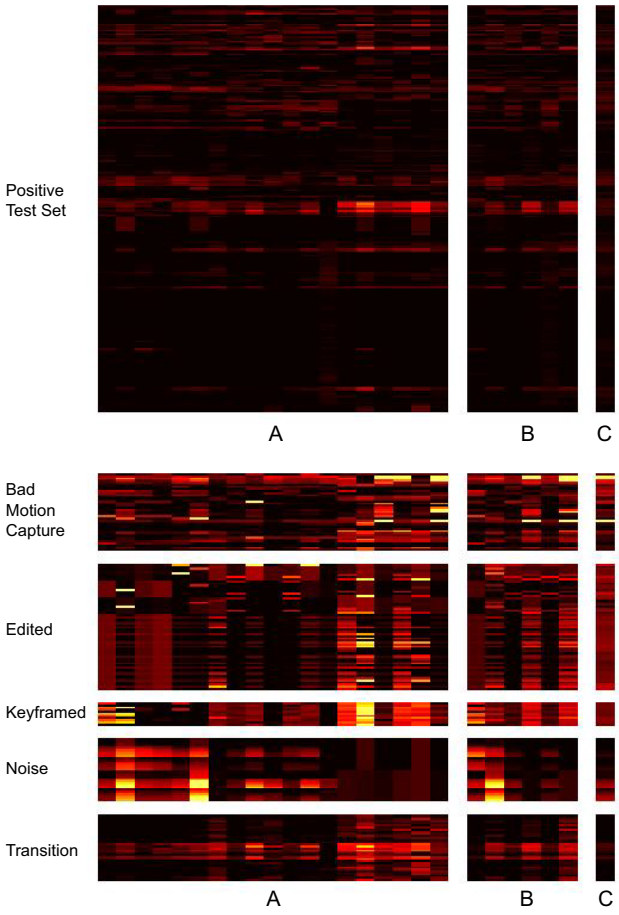


Figure 5: Response of the ensemble of HMM to the positive and the negative testing data. Each row shows the responses of all 26 models to a particular testing sequence. The intensity of the color (red to yellow) indicates a decreasing score (more unnatural). Each column corresponds to a single ensemble, grouped as follows: A-joints, B-limbs, and C-full-body, (see Figure 2).

to identify for all of the methods, with the exception of SLDS and ensemble of SLDS. During a bad transition, the velocities change due to blending in a way that is locally smooth, but is inconsistent with the dynamics of the initial and final motion. We hypothesize that the good performance of the SLDS models can be attributed to

their ability to correctly model longer-term temporal properties.

Table 1 also describes the number of parameters in each of the models. These parameters are the degrees of freedom that the model can exploit in fitting the data and provide a crude measure of the representational resources of the models. The ensemble of MoG and of SLDS have many more parameters than the ensemble of HMM but produce slightly inferior performance. This discrepancy is perhaps a sign that these more complex models may be overfitting the training data.

Figure 4 further explores the performance of the ensemble of HMM. Each type of model is shown in a different color: single joints, limbs, and full body. As expected, the ensemble model that is computed by combining the scores of the individual HMM has significantly better performance than any single HMM. The individual HMM are fairly tightly bunched indicating that each potentially has value in the computation of the overall score.

One advantage of the ensemble approach is that it can be used not only to detect unnatural motions but also to localize problem areas. This property is illustrated in Figure 5 where the color of a block indicates whether a particular HMM found each motion to be natural (black) or unnatural (red to yellow). In order to detect unnatural motion in an individual joint or limb, we compare the normalized score from the corresponding smaller model with the threshold that gives the equal error rate for the ensemble classifier. Joints that are below threshold are flagged as unnatural and rendered with a color that is proportional to the score. Two unnatural motions are visualized in Figure 1 with the joints that were detected as unnatural shown in red-yellow. By localizing problem areas to particular joints or limbs, we found errors in our previously published database that had not been noticed when the data was cleaned and processed.

Our user study produced a true positive rate of 93% and a 7% false positive rate. The subjects were drawn from a variety of disciplines and had not spent any significant time studying human motion data so it is perhaps not surprising that their classification did not agree completely with that of the authors when they assembled the testing database. Informal interviews with the subjects indicated that they were sometimes confused by the absence of objects that the character should have been interacting with (a box that was stepped onto, for example). If the semantics of the motion was not clear, they were likely to label it as unnatural. The subjects also missed some errors in the motion, most commonly those of short duration.

The training time for each of these statistical methods was significant, ranging from a few hours for the simpler methods to several days for the ensemble methods. The testing time is not long however, we were able to test the entire set of motions in 20 minutes.

## 6 Discussion

Our measures cannot be significantly better than the motion database of positive examples used to train them. Motions that are quite distant from those in the training set will likely be judged unnatural even if they are in fact natural. In our experiments, we have seen that unusual motions that have little in common with other motions in the database are sometimes labeled unnatural. For example, we have only a few examples of falling in the motion database and "natural" examples of that behavior were judged as unnatural by our measures. On the other hand, we have also seen evidence that the measures do generalize. For example, our testing set included walking while picking up a coffee mug from a table. This motion was judged natural by most of the methods although based on a visual inspection, the closest motions in the training dataset were a two arm reach while standing, walking, and sweeping with a broom.

Negative examples often bear the imprint of the algorithm used to create them. For example, carelessly edited motions might evidence unbalanced postures or foot sliding if inverse kinematics was not used to maintain the foot constraints. Similarly, motions that include bad transitions often have significant discontinuities in velocity as the blending routine attempts to smooth between two distant poses with differing velocities. We have attempted to address this concern by testing on a wide variety of common errors: motions that were aggressively edited in a commercial animation package, motions that were keyframed by an inexperienced animator, badly cleaned motion capture data, bad (and good) transitions, and motions with synthetic noise added. A larger variety of negative training examples would allow a more rigorous assessment of competing techniques.

Despite our attempt to span the space of motion errors with our negative testing set, other common errors may not be reliably detected. For example, our methods will likely not detect very short errors because the score on a motion is computed as an aggregate over an entire sequence of motion. The magnitude of the error caused by a single glitch in the motion will be reduced by the high percentage of good, natural motion in the sequence. This particular flaw does not seem serious, however, because a special-purpose detector could easily be created for glitch detection. Furthermore, most modern editing and synthesis techniques avoid this kind of error.

Our measures are also not very effective at detecting otherwise natural motion that has been slowed down by a factor of two. Such a slow-down is sometimes difficult for human observers to detect as well, particularly for behaviors that do not include a flight phase to provide decreased gravity as a reference. We believe that our methods do not perform well on these motions because the poses and lower velocities seen in these motions are "natural" in the sense that they would be seen in such natural behaviors as slow walks. Furthermore, the HMM have self-loops that allow slower motions to pass without significant penalty.

Apart from their use as an evaluation tool, measures of naturalness could be used to improve the performance of both motion synthesis and motion editing tools by identifying motion produced by those algorithms that was likely not natural. Those labels could be used to adjust the threshold of a particular transition metric, (for example, Wang and Bodenheimer [2003]) or to assess the value of a new editing algorithm.

In order to facilitate comparison between models, we used a standard approach to dimensionality reduction and standard constraints such as diagonal covariances to reduce the number of model parameters. In future work we plan to explore dimensionality reduction approaches for the SLDS model that exploit the dynamics of the data more effectively (for example, [Soatto et al. 2001]).

Our approach to measuring the naturalness of a motion via ensembles of smaller models was quite successful. However, it is likely that the methods could be improved, given that human observers perform significantly better on our test set. In the approach reported here, we used our knowledge about the synergies of human motion to pick appropriate feature groups but feature selection from among a larger set of features might produce better results. We combined the scores of the small models by normalizing and then simply picking the worst score. Other, more sophisticated, methods for normalizing or computing the score might provide better results.

In addition to screening for naturalness, our approach might work for screening for the style of a particular character. For example, a

measure could be trained on all the keyframe motion for a particular cartoon character. Each new motion sequence could then be tested against that measure to determine if the motions were "in character." If not, those motions could be flagged for closer inspection and perhaps re-animation by the animator.

**Acknowledgments:** The authors would like to thank Moshe Mahler for his help in modeling and rendering the images for this paper and Justin Macey for his assistance in collecting and cleaning the motion capture data. The authors would also like to thank Jackie Assa for using his method [Assa et al. 2005] to select the images in Figure 1 and Vladimir Pavlović for the use of his EPMT software for HMM and SLDS learning. This material is based upon work supported in part by the National Science Foundation under NSF Grants CNS-0196217, IIS-0205224, IIS-032622, IIS-0133779, and a Graduate Research Fellowship awarded to the second author. Alias/Wavefront donated their Maya software for use in this research.

## References

ARIKAN, O., AND FORSYTH, D. A. 2002. Interactive motion generation from examples. *ACM Transactions on Graphics 21(3)*, 483–490.

ASSA, J., CASPI, Y., AND COHEN-OR, D. 2005. Action synopsis: Pose selection and illustration. *ACM Transactions on Graphics 24(3)*.

BRAND, M., AND HERTZMANN, A. 2000. Style machines. In *Proceedings of ACM SIGGRAPH 2000*, 183–192.

COLE, R. A., 1996. Survey of the state of the art in human language technology. http://cslu.cse.ogi.edu/HLTsurvey.

FARID, H., AND LYU, S. 2003. Higher-order wavelet statistics and their application to digital forensics. In *IEEE Workshop on Statistical Analysis in Computer Vision (in conjunction with CVPR 2003)*.

GLEICHER, M. 2001. Comparing constraint-based motion editing methods. *Graphical Models 63(2)*, 107–134.

HAMID, R., JOHNSON, A., BATTA, S., BOBICK, A., ISBELL, C., AND COLEMAN, G. 2005. Detection and explanation of anomalous activities. In *IEEE Conference on Computer Vision and Pattern Recognition*. To appear.

HARA, K., OMORI, T., AND UENO, R. 2002. Detection of unusual human behavior in intelligent house. In *Neural Networks for Signal Processing XII-Proceedings of the 2002 IEEE Signal Processing Society Workshop*, 697–706.

HARRISON, J., RENSINK, R. A., AND VAN DE PANNE, M. 2004. Obscuring length changes during animated motion. *ACM Transactions on Graphics 23(3)*, 569–573.

IKEMOTO, L., AND FORSYTH, D. A. 2004. Enriching a motion collection by transplanting limbs. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 99–108.

KOVAR, L., AND GLEICHER, M. 2004. Automated extraction and parameterization of motions in large data sets. *ACM Transactions on Graphics 23(3)*, 559–568.

KOVAR, L., GLEICHER, M., AND PIGHIN, F. 2002. Motion graphs. *ACM Transactions on Graphics 21(3)*, 473–482.

LEE, J., CHAI, J., REITSMA, P., HODGINS, J., AND POLLARD, N. 2002. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics 21(3)*, 491–500.

LERNER, U. 2002. *Hybrid Bayesian Networks for Reasoning about Complex Systems*. PhD thesis, Stanford University.

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

O'SULLIVAN, C., DINGLIANA, J., GIANG, T., AND KAISER, M. K. 2003. Evaluating the visual fidelity of physically based animations. *ACM Transactions on Graphics 22(3)*, 527–536.

PAVLOVIĆ, V., REHG, J. M., AND MACCORMICK, J. 2000. Learning switching linear models of human motion. In *Proceedings of Advances in Neural Information Processing Systems (NIPS 2000)*, 981–987.

PERLIN, K. 1995. Real time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics 1(1)*, 5–15.

POLLICK, F., HALE, J. G., AND MCALEER, P. 2003. Visual perception of humanoid movement. In *Proceedings Third International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems 101*, 107–114.

RABINER, L. R., AND JUANG, B.-H. 1993. *Fundamentals of Speech Recognition*. Prentice Hall.

REITSMA, P. S. A., AND POLLARD, N. S. 2003. Perceptual metrics for character animation: Sensitivity to errors in ballistic motion. *ACM Transactions on Graphics 22(3)*, 537–542.

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

SOATTO, S., DORETTO, G., AND WU, Y. 2001. Dynamic textures. In *IEEE International Conference on Computer Vision*, vol. 2, 439–446.

SULEJMANPASIC, A., AND POPOVIC, J. 2005. Adaptation of performed ballistic motion. *ACM Transactions on Graphics 24(1)*, 165–179.

TROJE, N. K. 2002. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision 2*, 371–387.

VAN TREES, H. L. 1968. *Detection, Estimation, and Modulation Theory*, vol. 1. John Wiley.

VICON MOTION SYSTEMS, 2005. http://www.vicon.com/.

WANG, J., AND BODENHEIMER, B. 2003. An evaluation of a cost metric for selecting transitions between motion segments. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 232–238.

WANG, J., AND BODENHEIMER, B. 2004. Computing the duration of motion transitions: an empirical approach. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 335–344.

WILEY, D. J., AND HAHN, J. K. 1997. Interpolation synthesis of articulated figure motion. *IEEE Computer Graphics and Applications 17(6)*, 39–45.

ZHONG, H., SHI, J., AND VISONTAI, M. 2004. Detecting unusual activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 819–826.