Modeling Dynamics of Expressive Body Gestures In Dyadic Interactions

Zhaojun Yang, Student Member, IEEE, and Shrikanth Narayanan, Fellow, IEEE

Abstract—Body gestures are an important non-verbal expression channel during affective communication. They convey human attitudes and emotions as they dynamically unfold during an interpersonal interaction. Hence, it is highly desirable to understand the dynamics of body gestures associated with emotion expression in human interactions. We present a statistical framework for robustly modeling the dynamics of body gestures in dyadic interactions. Our framework is based on high-level semantic gesture patterns and consists of three components. First, we construct a universal background model (UBM) using Gaussian mixture modeling (GMM) to represent subject-independent gesture variability. Next, we describe each gesture sequence as a concatenation of semantic gesture patterns which are derived from a parallel HMM structure. Then, we probabilistically compare the segments of each gesture sequence extracted from the second step with the UBM obtained from the first step, in order to select highly probabilistic gesture patterns for the sequence. The dynamics of each gesture sequence are represented by a statistical variation profile computed from the selected patterns, and are further described in a well-defined kernel space. This framework is compared with three baseline models and is evaluated in emotion recognition experiments, i.e., recognizing the overall emotional state of a participant in a dyadic interaction from the gesture dynamics. The recognition performance demonstrates the superiority of the proposed framework over the baseline models. The analysis of the relationship between the emotion recognition performance and the number of the selected segments also indicates that a few local salient events, rather than the whole gesture sequence, are sufficiently informative to trigger the human summarization of their overall global emotion perception.

Index Terms—Body gesture, gesture patterns, gesture dynamics, emotion recognition, motion capture.

1 INTRODUCTION

In human communication, body gestures are an essential element of non-verbal behavior to express interpersonal attitudes, feelings and affect [1] [2] [3]. Research on recognizing emotions using body gesture expressions has hence received much interest [4] [5]. Existing research has mostly focused on specific types of short-term body gestures, e.g., knocking, walking or waving, suggesting that body gestures are emotion-specific to some extent [6] [7]. A quantitative knowledge of how body gestures are involved in emotion communication as they dynamically unfold during an interpersonal interaction is however still largely understudied. Understanding the dynamics of body gestures in the context of emotion expression can have a significant impact on automatic emotion recognition as well as the design of advanced human-machine interfaces. For example, the pedagogical agent developed in [8] incorporates a probabilistic model of a user’s affect. By monitoring the user’s affect in the interactions during educational games, the agent adjusts the decisions about generating appropriate interventions to improve the user’s learning effectiveness.

The major challenge in modeling dynamics of body gestures is the high degree of gesture variability. Human body gestures are complex in nature in terms of both temporal patterning and spatial details, varying both within and across individuals as well as over different time scales. The gestures produced during interpersonal interactions are even more intricate due to the interaction-context factors such as the stances assumed, the variety and variation in communication intentions and the behavior of conversational partners. However, analogous to the compositional view of visemes in lip motion, there are elementary patterns that have been defined for body gestures, i.e., gesture phrases/units. In the gesture model proposed by Kendon [9], a gesture phrase defines the basic gesture element and a natural continuous gesture can be decomposed into multiple gesture phrases. There have been a few approaches for representing gesture dynamics [5] [10] [11] including the semantic pattern based description of body gestures which provides us the basis of modeling gesture dynamics in this paper as well. These existing works however are mainly designed for short-term body gestures with rhythmically repeating patterns. Moreover, they only consider variations locally within each individual gesture sequence, which may induce subject-dependent characteristics and undermine the essential dynamic cues. Hence, robustly representing gesture dynamics, especially for long-term gesture sequences with high dynamical complexity, becomes another pressing need and is studied in this work.

The main objective of this work is to robustly model the dynamics of expressive body gestures occurring in long-term dyadic interactions. We propose a statistical framework based on the high-level semantic gesture patterns. In this framework, we first construct a universal background model (UBM) using Gaussian mixture model (GMM) to describe the global subject-independent variability of gesture features. Next, we employ a parallel HMM structure to segment gesture sequences and extract semantic gesture patterns. Further, we probabilistically fit the segments of each gesture sequence with the UBM, in order to select...
highly probabilistic patterns for the sequence. The gesture dynamics of a sequence are represented as a statistical variation profile computed from the selected salient segments, and are described in a well-defined kernel space.

The advantages of our framework for robust gesture dynamics modeling are summarized as follows: 1) To address the high variability and complex structure of body gestures, each gesture sequence is described based on semantic gesture patterns. 2) To unify the information of all the gesture sequences, the gesture patterns of each sequence are fitted to a statistical UBM. This procedure helps to remove individual idiosyncrasies within a gesture sequence. 3) To capture the relationship between the modeled dynamics of two gesture sequences, we propose a kernel function based approach.

Our work is evaluated on the freely-available multimodal USC CreativeIT database that consists of goal-driven improvised interactions [12] [13]. It contains detailed full body Motion Capture (MoCap) data, providing a rich resource for studying body gestures during expressive interactions. We focus on hand gesture and head motion which are expressive and informative in communication. Our model is evaluated on emotion recognition tasks. The experimental results show that the proposed UBM-based model outperforms other examined approaches in terms of recognition accuracy. We also observe that hand gesture carries more affect-specific information than rigid head motion.

The rest of the paper is organized as follows. We discuss related work on the gesture-affect relationship, especially on body gesture modeling for emotion recognition, in Section 2. The proposed framework for representing gesture dynamics is presented in Section 3. We introduce the multimodal CreativeIT database, the gesture feature extraction and emotion annotation in Section 4. We present the analysis of the extracted gesture patterns in Section 5. The baseline models are described in Section 6, followed by the experimental results of emotion recognition tasks in Section 7. This paper’s conclusions and future work are given in Section 8.

2 Related Work

Body gestures are used as an integral means of expression in human communication. There is an extensive literature studying the gesture-emotion relationship and showing body gestures are as important as facial expressions in emotion conveyance. Wallbott analyzed the emotional content of acted body movements and postures using coding schemata, showing that body movements and body posture are indicative of specific emotions [6]. De Meijer also corroborated that different emotion categories can be inferred from the intensity and the types of body movements [7]. Mehrabian and Friar found that body orientation is an important indicator of the communication attitude of a participant towards one’s interlocutor [14]. Researchers have also demonstrated that gestures are communicatively intended by speakers and express the underlying cognitive architecture in a conversation [15] [16].

The above-mentioned psychological studies on the gesture-emotion relation have inspired work on automatic emotion recognition from body expressions. Early work in this direction has focused on detecting emotions from acted and stylized body movements and postures. For example, gait patterns have been analyzed with respect to the affective state of an individual in both categorical and dimensional emotion spaces [17]. Kapur et al. investigated low-level physical features of stylized body movements, e.g., velocities and accelerations of marker positions, for emotion discrimination. The low-level dynamics have shown to be effective in distinguishing the four basic emotions of sadness, joy, anger and fear [18].

Since body gestures commonly occur in daily human communication, research efforts have also been devoted to exploring the potential of using interaction gestures for emotion recognition. Metallinou et al. have used body language information for automatically tracking the continuous emotional attributes of activation, valence and dominance of a participant over affective communication [19]. They describe the body language of a participant in terms of body movements and postures, such as hand velocities, head angles and body positions. Nicolaou et al. tracked the continuous human affect during a human-agent conversation by fusing shoulder movements, facial expressions and speech cues [20]. To consider the temporal dynamics of non-acted body gestures, a Recurrent Neural Network algorithm was employed for emotion recognition in the context of a video game [21].

In spite of the success of the low-level gesture dynamics for emotion detection, such descriptions are insufficient to capture the structure and dynamical cues of natural body gestures produced in long-term interactions, due to the sophisticated multi-scale nature of human gestures. To model the complex structure of body gestures, researchers have hence attempted to decompose the complex motion into simple isolated elements. Bernhardt and Robinson represented the dynamical cues of the knocking motion using motion primitives that are derived from K-means clustering [5]. For robust emotion recognition, the primitive-based motion dynamics are modeled in an individual-unbiased way by removing subject-specific characteristics. They further extended this framework for detecting emotions from natural action sequences [22]. Camurri et al. analyzed the affective states of dance sequences by developing mid-level features from the segmentation of the dance trajectories [23].

In order to obtain meaningful motion patterns, several approaches have been proposed for learning the primitives of human actions from motion sequences. For example, Levine et al. derived gesture subunits from motion data by detecting zero points of the angular velocity [24]. Based on these motion segments, they further applied Ward hierarchical clustering to identify the recurring motion subunits. A probabilistic PCA based algorithm has been proposed in [25] to segment motion data into distinct actions, under the assumption that a motion transition occurs when the distribution of the motion data changes. Zhou et al. proposed an unsupervised hierarchical framework that combines the kernel k-means and a generalized dynamic time alignment kernel, for temporally segmenting and clustering multi-dimensional time series [26]. This framework has shown promising results in clustering a small number of human actions. However, like most kernel methods, this approach also suffers from the high computation and storage complexity on large-scale data, due to the computation of an
In this work, our framework for body gesture modeling is based on the elementary gesture patterns identified from the parallel HMM model. In order to minimize the individual-dependent variations and to unify the information from all the gesture sequences, we construct a GMM-UBM model to describe the global gesture variability. Instead of representing dynamics locally within each sequence as most of the above-mentioned studies did, we align the patterns of each sequence with the subject-independent GMM-UBM model for robustly representing the dynamics. This GMM-UBM model has been widely applied for acoustic feature modeling in text-independent speaker verification [29][30]. It captures inter-speaker variability and is viewed as a viable speaker model. Li et al. developed a GMM-UBM based face matching system for pose variant face verification [31]. Liu et al. proposed a framework also based on GMM-UBM for dynamical face recognition [32], which is similar in spirit to the approach adopted in the present paper. In their work, each video clip is described by uniformly sampled cuboids or segments. In our work, the gesture patterns are automatically identified, which provides semantic meanings as well as robust gesture descriptions. In addition, we consider more comprehensive gesture statistics for representing the dynamics. The relationship between the representations of dynamics is further described in a well-defined kernel function.

3 Framework Overview

Fig. 1 illustrates an overview of the framework that we propose for body gesture modeling. In the framework, a gesture sequence $S$ is represented by the gesture features $\{f_1, f_2, \cdots, f_T\}$, where $T$ is the length of $S$ and $f_t$ is a gesture feature vector at time $t$. First, we build a UBM to statistically describe the global variability profile of gesture features from different individuals. A GMM is employed to learn this background model. Each component of the learned GMM represents one universal variation mode of the gesture features at the frame level. We further use the individual gesture sequences to train a parallel HMM structure for identifying temporally recurring gesture patterns. Henceforth, each gesture sequence is partitioned to short segments corresponding to the gesture patterns. Finally, the dynamics of an entire gesture sequence are modeled by a statistical variation profile that is computed from the statistically salient segments with respect to each component of the UBM.

In the rest of this section, we present the construction of the background model of gesture features using GMM (Section 3.1), the details of gesture clustering using the parallel HMM structure (Section 3.2) and the description of the statistical gesture dynamics modeling (Section 3.3).

3.1 Universal Background Model Construction

To statistically describe the global variability profile of the gesture features from different subjects, we build a universal background model to represent the subject-independent distribution of gesture features. In this work, we use GMMs to learn the background model. GMMs have shown great success for universal background modeling in speaker verification [29][30] and facial expression recognition [31][32]. Each component of the GMM describes one class of universal variations of gesture features at the frame level. A GMM background model, $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$, is constructed based on the feature vectors of gesture sequences $\{f_1, f_2, \cdots, f_T\}$,

$$p(f|\Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(f; \mu_k, \Sigma_k),$$

where $K$ is the number of mixtures, $\pi_k$, $\mu_k$ and $\Sigma_k$ are the weight, mean vector and covariance matrix of the $k$-th component, and $\sum_{k=1}^{K} \pi_k = 1$. In this work, $\{\Sigma_k\}_{k=1}^K$ are specified as diagonal covariance matrices. The GMM parameters $\Theta$ can be estimated based on the maximum likelihood criterion using Expectation Maximization (EM).

3.2 Extraction of Gesture Patterns

The elementary gesture patterns have not been well established quantitatively so far, due to the nature of the gesture structure — the high degree of variability across persons and contexts as well as along different temporal scales. In this work, we hence identify the recurring patterns in an unsupervised manner. We employ the parallel HMM model for the extraction of elementary phrases [27]. This model provides flexibility and efficiency in modeling the variations in the structure and durations of gesture patterns.

The parallel HMM model, $\Lambda$, is composed of $M$ parallel left-to-right HMMs $\{\lambda_m\}_{m=1}^M$, where each branch $\lambda_m$ has $Q$ states $\{s_{m,1}, s_{m,2}, \cdots, s_{m,Q}\}$, as shown in Fig. 2. $M$
corresponds to the number of clusters, i.e., the number of gesture patterns. All the branches share the same starting and ending states $s_s$ and $s_e$ to ensure the continuity at the boundaries between segments. We empirically select the number of states in each branch of the HMM model $\Lambda$ as $Q = 10$, corresponding to the minimum gesture pattern duration of 10 frames ($\frac{1}{6}$ sec assuming 60 frames/sec). The sequence of gesture vectors $S$ is used to train the model $\Lambda$. The segmentation and clustering are performed by maximizing the likelihood using Viterbi decoding:

$$\{\epsilon_l, m_l\}_{l=1}^L = \arg \max_{\{\epsilon_l, m_l\}} \prod_{l=1}^L P(\epsilon_l | \lambda_{m_l}), \quad (2)$$

where $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_L\}$ are the $L$ number of gesture segments produced by the model $\Lambda$. Each segment $\epsilon_l$ which is represented by a feature set, $F_{\epsilon_l} = \{f_{t_l}, f_{t_l+1}, \ldots, f_{t_{l+1}-1}\}$, is assigned to one of the $M$ clusters with a cluster label $m_l$ $(m_l \in \{1, 2, \ldots, M\})$. The segments $\epsilon_l$ associated with cluster labels $m_l$ represent the recurring gesture patterns that are captured by the probabilistic structure $\Lambda$. We accordingly define the frame-level labels $g_l$ based on this association, i.e., $g_{l} = m_l$, if $f_{l} \in \epsilon_l$. As a result, a sequence of gesture vectors $S$ can be mapped into a sequence of cluster labels $g = \{g_1, g_2, \ldots, g_T\}$.

### 3.3 Statistical Gesture Dynamics Modeling

The background model $\Theta$ obtained in Section 3.1 represents the global subject-independent distribution of gesture features, and the gesture patterns extracted in Section 3.2 provide a high-level semantic description of a gesture sequence. Given any gesture sequence, the probabilistic fitting of its gesture patterns to the background model could unify the information from all the sequences and provide robust dynamics characterization by removing subject-specific variations within the sequence. We therefore formulate the modeling of statistical gesture dynamics as follows.

As described in Section 3.2, a gesture sequence $S$ is partitioned into short segments $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_L\}$, where $\epsilon_l$ is represented by a feature set $F_{\epsilon_l} = \{f_{t_l}, f_{t_l+1}, \ldots, f_{t_{l+1}-1}\}$. From the $k$-th component in the background model, the probability, $p_{\epsilon_l}^k$, of $\epsilon_l$ can be computed as,

$$p_{\epsilon_l}^k = \frac{1}{t_{l+1} - t_l} \sum_{t=t_l}^{t_{l+1}-1} \pi_k N(f_t; \mu_k, \Sigma_k). \quad (3)$$

A segment with a higher probability $p_{\epsilon_l}^k$ is statistically salient within the sequence $S$. This salient gesture pattern reflects a subject’s central gesture expression which may play an important role in conveying one’s internal mental state. Moreover, a salient pattern stands out through a
We further define the distance between sets of covariance descriptors,\(\{\text{Cov}^k\}_{k=1}^K\), as the sum of individual covariance distances,

\[
dist_{\text{cov}}(S_i, S_j) = \sum_{k=1}^K \text{dist}(\text{Cov}^k_i, \text{Cov}^k_j).
\]

(7)

Accordingly, we define the distance between sets of probability descriptors of \(S_i\) and \(S_j\) using \(L_2\) norm,

\[
dist_{\text{prob}}(S_i, S_j) = \sum_{k=1}^K \|p^k_i - p^k_j\|_2.
\]

(8)

Both types of distance metrics in Eq. (7) and (8) measure the distance between statistical gesture patterns of a sequence, i.e., characterizing the long-term statistical gesture changes. In addition, the associated probability set \(p^k\) captures the general likelihood of the salient patterns in \(S\). The dynamics of a gesture sequence can therefore be represented by a statistical profile that is suited with both covariance and probability sets, \(\{\text{Cov}^k\}_{k=1}^K\) and \(\{p^k\}_{k=1}^K\).

The covariance descriptors, \(\{\text{Cov}^k\}_{k=1}^K\), are symmetric positive definite (SPD) matrices which lie on a Riemannian manifold. We exploit the Log-Euclidean distance between any two SPD matrices, \(\text{Cov} = U\Sigma U^T\), the matrix logarithm can be computed as,

\[
\log(\text{Cov}) = U \log(\Sigma) U^T.
\]

(6)

We further define the distance between sets of covariance descriptors of sequences \(S_i\) and \(S_j\) as the sum of individual covariance distances,

\[
dist_{\text{cov}}(S_i, S_j) = \sum_{k=1}^K \text{dist}(\text{Cov}^k_i, \text{Cov}^k_j).
\]

(7)

4.1 Gesture Feature Extraction

After capturing the motion data, we manually mapped the 3D locations of the markers to the angles of different human body joints using MotionBuilder [35]. The mapped angles are then used as body gesture features. The joint angles are preferred instead of 3D coordinates to describe gestures, because they are more suitable for animation purposes [24] [27] and the subject-dependent gesture characteristics (e.g., the arm length) have been removed through the mapping process. In this work, we focus on two types of body gestures which are expressive and emotion-informative in communication: hand gesture and head motion [36]. Figure 4 illustrates the Euler angles of the hand (arm and forearm) and head joints in the \(x, y\) and \(z\) directions.

To incorporate the temporal dynamics, we augment the gesture features with their 1st order derivatives. The gesture feature vector \(f^i_t\) of the joint \(n\) at frame \(t\) is:

\[
f^i_t = [\theta^i_n, \phi^i_n, \psi^i_n, \Delta\theta^i_n, \Delta\phi^i_n, \Delta\psi^i_n]^T.
\]

(10)
\( \Delta \phi^a_t \) and \( \Delta \psi^a_t \) are their corresponding 1st order derivatives. The hand gesture features include the information of the four hand joints, \( i.e., \) left and right arms, as well as left and right forearms. The head gesture is then represented by \( \{ f_{\text{leftarm}}^a, f_{\text{rightarm}}^a, f_{\text{leftforearm}}^a, f_{\text{rightforearm}}^a \} \). Similarly, the head motion is related to only one joint, and is represented by a 6-D feature vector \( \mathbf{f}_{\text{head}} \).

**5 Analysis of Gesture Patterns**

An essential component of our framework is to extract meaningful gesture patterns as described in Section 3.2. In this section, we investigate the effect of the number of gesture patterns, \( M \), on the dynamics modeling. We then verify the validity of the derived clusters for representing semantic gesture patterns.

### 5.1 Segmentation of Gesture Patterns

The number of parallel HMMs \( M \), \( i.e., \) the number of gesture patterns, is an important parameter which affects gesture dynamics modeling. A small \( M \) provides a coarse-grained gesture description, while a large \( M \) leads to a noisy gesture representation. In order to identify a suitable number of gesture patterns, we employ a bigram model to capture the dynamical evolution of a gesture sequence with respect to different cluster numbers. A high-quality bigram model indicates an appropriate \( M \).

A bigram model is a first-order Markov model, popular in modeling word sequences (here sequences of gesture labels \( g \)) in language processing. We use the sequences \( g \) obtained in Section 3.2 to calculate the transition (bigram) probabilities of the gesture labels within a sequence. Given \( g_t \) at time \( t \) and \( g_{t-1} \) at previous time \( t-1 \), \( P(g_t | g_{t-1}) \) defines the bigram probability that the gesture \( g_t \) occurs if the previous gesture \( g_{t-1} \) has been observed. We use perplexity to evaluate the computed gesture bigram model. Perplexity is a popular measure to evaluate language (word sequence) models [40]. It quantifies the confusion of the current state, \( i.e., \) the average number of possible successors, from an information theoretic perspective. A lower perplexity indicates a better bigram model, and a higher perplexity implies more randomness in the derived sequences of gesture labels. The perplexity \( \text{ppl} \) of a bigram model is defined as:

\[
\text{ppl} = P(g)^{- \frac{1}{|g|}},
\]

where \( |g| \) is the length of a gesture sequence. The probability \( P(g) \) is computed using the bigram model as:

\[
P(g) = P(g_1) \prod_{t=2}^{|g|} P(g_t | g_{t-1}).
\]

However, this measure depends on the vocabulary size (the number of clusters \( M \)), \( i.e., \) a larger \( M \) leads to a higher perplexity. To alleviate this dependency, we apply...
The relationship for head motion (the circle marker). We can see that the normalized perplexity of the bigram model of hand gesture drops rapidly with an increasing number of clusters. The decreasing rate of head motion falls off from \( M = 50 \) for hand gesture slows down from \( M = 50 \) varying with the number of clusters. The decreasing rate of head motion \( M \). We mainly consider the top six clusters which contain the most number of gesture segments (The average number of segments in each of the top six clusters is around 350, while the mean segment number in each of the remaining clusters is around 80). To this end, we first extract statistical functionals of each gesture feature within a segment, e.g., mean, standard deviation, maximum, minimum, median, range, kurtosis and skewness. Hence, the variations of gesture segments are represented by vectors of gesture statistics. We further apply the parametric t-SNE, an unsupervised dimensionality reduction technique, to map the high-dimensional gesture variation space to a 2-dimensional latent space. The parametric t-SNE learns the parametric mapping by optimally preserving the local data structure in the low-dimensional latent space [42]. Fig. 7 visualizes the 2-D representations of the segment-level gesture dynamics in the top 6 clusters with respect to hand gesture and head motion. In general, the variations of gesture segments in different clusters are clearly separated. Such distinguishability between distinct clusters visually verifies the validity of the derived clusters for representing semantic gesture patterns.

In addition, we manually examined the gesture segments (longer than 1 sec) in the major six clusters with respect to hand gesture and head motion by watching the video clips. The frequent hand gesture patterns include moving both hands to the front, pointing with one’s left hand, pointing with one’s right hand, crossing arms in front of chest, opening arms and waving arms by the side. The common head motion patterns are raising one’s head, lowering one’s head, keeping still, turning to the left slightly, turning to the right slightly, and turning to the left mightly. Some of these patterns are congruent with the emotionally expressive gestures identified in the social psychology studies [6] [7]. For example, Wallbott found that the lateralized hand movements are frequent during the anger emotion and moving the head downward is the most typical for the disgust emotion [6].

![Fig. 6](image_url)

**Fig. 6.** Normalized perplexity of the bigram model of hand gesture or head motion varying with the number of clusters. The decreasing rate for hand gesture slows down from \( M = 50 \), and the changing rate for head motion falls off from \( M = 40 \).

![Fig. 7](image_url)

**Fig. 7.** Visualization of the dynamics of gesture segments of top six clusters in the 2-dimensional latent space for hand gesture (left) and head motion (right). Each circle represents the dynamics of each gesture segment, and colors of circles indicate the corresponding clusters.

5.2 Visualization of Gesture Patterns

As described in Section 3.2, the gesture patterns are extracted in an unsupervised clustering manner. Herein, we aim at validating the effectiveness of the derived clusters for representing the elementary gesture patterns. Since some extracted patterns could be as short as \( \frac{1}{2} \) sec (see Section 3.2), it is impossible for human observers to manually validate by watching these video clips. To approach this problem in a more systematic way, we visualize the variations of gesture segments in each cluster in a low-dimensional space. We mainly consider the top six clusters which contain the most number of gesture segments (The average number of segments in each of the top six clusters is around 350, while the mean segment number in each of the remaining clusters is around 80). To this end, we first extract statistical functionals of each gesture feature within a segment, e.g., mean, standard deviation, maximum, minimum, median, range, kurtosis and skewness. Hence, the variations of gesture segments are represented by vectors of gesture statistics. We further apply the parametric t-SNE, an unsupervised dimensionality reduction technique, to map the high-dimensional gesture variation space to a 2-dimensional latent space. The parametric t-SNE learns the parametric mapping by optimally preserving the local data structure in the low-dimensional latent space [42]. Fig. 7 visualizes the 2-D representations of the segment-level gesture dynamics in the top 6 clusters with respect to hand gesture and head motion. In general, the variations of gesture segments in different clusters are clearly separated. Such distinguishability between distinct clusters visually verifies the validity of the derived clusters for representing semantic gesture patterns.

In addition, we manually examined the gesture segments (longer than 1 sec) in the major six clusters with respect to hand gesture and head motion by watching the video clips. The frequent hand gesture patterns include moving both hands to the front, pointing with one’s left hand, pointing with one’s right hand, crossing arms in front of chest, opening arms and waving arms by the side. The common head motion patterns are raising one’s head, lowering one’s head, keeping still, turning to the left slightly, turning to the right slightly, and turning to the left mightly. Some of these patterns are congruent with the emotionally expressive gestures identified in the social psychology studies [6] [7]. For example, Wallbott found that the lateralized hand movements are frequent during the anger emotion and moving the head downward is the most typical for the disgust emotion [6].

6 BASELINE MODELS

We evaluate our UBM-based method for gesture dynamics modeling by comparing with three baselines on automatic emotion recognition tasks. Two of the baselines are also developed based on the gesture patterns from Section 3.2.

6.1 Low-level Physical Dynamics

The relationship between low-level physical dynamic cues and emotions has been extensively investigated by re-
such low-level dynamics have demonstrated the effectiveness for emotion recognition in some simple scenarios [18] [36] [43] [44]. A gesture feature vector \( f \) in our work (see Section 4.1) is represented by a series of 3D joint angles and their 1st order derivatives (velocities). To compute the low-level dynamics, we also include their 2nd order derivatives to describe acceleration. At the gesture sequence level, we consider mean, standard deviation, kurtosis, and skewness of the joint angles, velocities and accelerations.

### 6.2 Markov-based Dynamics

In Section 5.1, we apply the bigram (“language”) model to evaluate how the transition dynamics of gesture sequences depend on the number of clusters. Herein, we employ these Markov-based evolution dynamics as one of our baselines. Besides the bigram dynamics, we compute the unigram from each sequence of gesture labels. A unigram describes a probability (or frequency) distribution of clusters within a sequence, and a bigram captures the local dependency between adjacent gesture events. We further reduce the dimensionality of the bigram features using Principal Component Analysis (PCA) by preserving 90% of the total variance. Such Markov-based dynamics have also been explored for attitude recognition [10].

### 6.3 Graph-based Dynamics

A graph-based framework for modeling gesture dynamics has been proposed in [11]. In this framework, an undirected graph is constructed for each sequence with each graph node representing a gesture segment. The graph Fourier transform (GFT) is subsequently applied to produce gesture variability within each sequence. Similarly to the classic Fourier transform, the graph-based description is represented in different frequencies. Low-frequency representations describe the smoothness of a gesture sequence, whereas the high-frequency ones capture the oscillations. Frequencies are further grouped into low- and high-frequency subbands. Similar statistical functionals, such as mean, median, maximum, or minimum, are extracted from the graph-based representations in each subband. Such statistical features in the low- and high-frequency subbands define the graph-based dynamics. As shown in [11], compared to Markov-based dynamics, Graph-based measures could better capture long-term gesture variability.

### 7 Emotion Recognition Experiments

In this section, we evaluate our method on emotion recognition tasks. We conduct two experiments: intra-subject emotion evaluation — to classify the emotion label (see Section 4.2) of an actor over an interaction using one’s own gesture information (hand or head gesture); and inter-subject emotion evaluation — to classify the emotion label of an actor using the gesture information of one’s interaction partner. The first experiment focuses on showing the effectiveness of our method for gesture dynamics modeling; and the second one aims at demonstrating the complementary nature of the cross-subject gesture dynamics captured in the UBM-based approach. We use the leave-one-actor-out scheme, i.e., 16-fold cross validation since there are 16 actors in the database (see Section 4). We report the performance averaged over all the folds. The classification experiments are performed using the multi-class SVM classifier with the defined kernel in Eq. (9) for the UBM-based model and with an RBF kernel for the three baselines. In each of the 16 folds, the parameters of the SVMs are tuned exclusively on the training set by leaving one actor out. Specifically, the soft margin parameter \( c \) of the SVMs is evaluated at 0.0001, 0.001, 0.01, 0.1 and 1. The scale parameters of the kernels, i.e., \( \sigma_{cov} \) and \( \sigma_{p} \) in the UBM-based method and \( \sigma \) in the baseline methods, are tuned as \( 2^q \), where \( q \in \{-5,-4, \ldots, 4, 5\} \). In the Markov-based method, we investigated the performance using unigram, bigram and their combination. The unigram features always perform the best. Similarly in the Graph-based approach, the performance using low- and high-frequency descriptions as well as their combination was examined. The best results were achieved by the high-frequency features. In addition, we examine the performance of the UBM-based method respectively using the parallel HMM structure and the aligned component analysis (ACA) [26] for the extraction of gesture patterns. In order to perform the ACA approach on the long-term gesture sequences, we reduce the number of frames in each sequence following the technique in [26]. To be consistent with the HMM structure, we use 50 hand gesture patterns and 40 head motion patterns in the ACA approach. The maximum length of a segment in ACA is set to 200 which is also the maximum segment length obtained using the HMM structure.

#### 7.1 Experimental Results

**Intra-Subject Emotion Recognition**

Table 1 presents the results of classifying the emotion label of a participant into three and four emotional clusters using one’s own hand gesture over an interaction. For the 3-cluster classification, we obtain the best baseline accuracy of 55.5%.
when using the low-level dynamic features. Our method (HMMs) improves the performance to 63.7%. For the 4-cluster classification, the performance has been upgraded to 48.9% by the UBM-based model (HMMs), from the best baseline result of 45.4% with the Graph-based method. Similar results using head motion are shown in Table 2. The UBM-based model (HMMs) achieves the best performance of 61.0% and 50.5% for the 3-cluster and 4-cluster classification respectively.

The experimental results first show that the statistical dynamics derived by the UBM-based model outperform the baselines in all the cases. Our framework explicitly selects statistically salient patterns by fitting each sequence to the subject-independent variability model. Therefore, the statistical variation profile that is computed among the salient patterns could robustly characterize gesture variations by excluding individual idiosyncrasies and concentrating essential dynamics within a sequence. Furthermore, hand gesture generally exhibits a higher discriminative power in distinguishing distinct emotions, compared to rigid head motion. This may be due to a greater degree of hand gesture variability as analyzed in Section 5.1. The richer expressiveness of hand gesture provides more information for discriminating distinct emotion categories. In contrast, the lower degree of head motion variability could restrict the emotion expression to some extent. Compared to hand gesture, it is more likely that the same head motion is used for expressing different emotions, which may bring confusions for emotion discrimination. We can also observe that Graph-based descriptors mostly outperform the other two baseline features using either hand gesture or head motion, which may result from the benefits that the Graph-based model could better capture the long-term sequence variations. Note that the Markov-based and Graph-based baselines as well as the UBM-based model are all based on the high-level gesture patterns derived in Section 3.2. They generally exceed the low-level physical dynamical cues in terms of the recognition performance. One implication might be that the low-level physical dynamics are not sufficient to capture the great spatial-temporal variability of human gestures, especially in a long-term interpersonal interaction.

**Inter-subject Emotion Recognition**

Table 3 and 4 present the results of classifying the emotion label of a subject using hand gesture or head motion of one’s conversational partner. Generally, we can observe a degraded recognition performance using cross-subject cues, compared to using one’s own. However, all the methods show certain level of effectiveness in cross-subject emotion recognition. This indicates the complementary nature of cross-subject gesture behavior, i.e., the gesture behavior of an interaction individual provides information about the emotional state of the other to some extent due to the inherent coupling of dyad’s mental and cognitive states during an interaction. In this experiment, our method still outperforms the baselines in all the cases. For example, it (HMMs) achieves the best accuracy of 60.6% and 48.8% using hand gesture respectively in the 3-cluster and 4-cluster classification. This improvement suggests that the gesture dynamics from the UBM-based model are informative of both intra-subject and cross-subject emotional states.

It is interesting to observe that the baseline models in some cases achieve even higher performance of inter-subject emotion recognition compared to that in the intra-subject tasks. On the one hand, the better performance in the inter-subject tasks supports the existence of interpersonal coordination of body gestures during dyadic interactions. On the other hand, this observation may also reveal weakness in the baseline approaches for body gesture modeling. The baseline representations contain both intra-subject and cross-subject characteristics simultaneously. For example, the same low-level physical dynamics are applied in each emotion recognition task. As a result, the two types of information are mutually influenced such that the modeled dynamics are even more informative regarding the interlocutor’s state than regarding one’s own. In contrast, the UBM-based method selects gesture patterns separately with respect to each evaluation task. The task-specific selection procedure may help disentangle the intra-subject and cross-subject factors, resulting in task-informative dynamical representations.

In both intra-subject and inter-subject experiments, the performance of the UBM-based method using ACA is lower, compared to using the HMMs structure. This result implicitly demonstrates the effectiveness of the parallel HMM model for extracting semantic gesture patterns. In contrast to the parallel HMM model, ACA performs segmentation and clustering respectively for each individual sequence, which may generate sequence-specific rather than use-generic gesture segments. Moreover, some important dynamical cues may be removed in the down-sampling process before applying ACA.

### Table 3

<table>
<thead>
<tr>
<th>Method</th>
<th>C = 3</th>
<th>C = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Low-level</td>
<td>47.6</td>
</tr>
<tr>
<td></td>
<td>Markov-based</td>
<td>45.4</td>
</tr>
<tr>
<td></td>
<td>Graph-based</td>
<td>57.3</td>
</tr>
<tr>
<td>UBM-based</td>
<td>ACA</td>
<td>57.6 (N = 7)</td>
</tr>
<tr>
<td></td>
<td>HMMs</td>
<td><strong>60.6 (N = 5)</strong></td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Method</th>
<th>C = 3</th>
<th>C = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Low-level</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>Markov-based</td>
<td>44.3</td>
</tr>
<tr>
<td></td>
<td>Graph-based</td>
<td>56.6</td>
</tr>
<tr>
<td>UBM-based</td>
<td>ACA</td>
<td>56.5 (N = 3)</td>
</tr>
<tr>
<td></td>
<td>HMMs</td>
<td><strong>57.7 (N = 5)</strong></td>
</tr>
</tbody>
</table>
7.2 The Effect of Salient Gesture Patterns

In the UBM-based framework, there is one key parameter impacting gesture dynamics modeling, i.e., the number of salient gesture segments, $N$, chosen for dynamics construction. Investigating the influence of $N$ on conveying the emotional state of an individual could help us to understand the way that human annotators summarize the local perceived gesture events to produce an overall emotional judgement about an interaction. The study of the perception mechanism of annotators is essential for behavioral science where human assessment is the main approach for various research analyses [45] [46]. Herein, we examine how the emotion recognition performance is related to $N$.

Fig. 8 presents the mean intra-subject recognition accuracy in relation to the number of selected segments $N$, using hand gesture and head motion respectively. We can observe that the performance generally increases in the beginning and then decreases as $N$ rises. A better performance is usually achieved when $N$ is around 15. This changing trend indicates that a few local salient events, rather than the entire gesture sequence, are sufficiently informative to trigger the human summarization of the global emotion perception.

Fig. 9 presents the mean inter-subject recognition accuracy in relation to the number of selected segments $N$, using hand gesture and head motion respectively. In contrast to Fig. 8, the inter-subject performance generally increases more rapidly in the beginning and starts falling from a relatively smaller value of $N$. The best performance is usually achieved at around $N = 5$. It is interesting to observe that fewer salient segments are needed to summarize the global rating of the cross-subject emotion, compared to those used for abstracting the intra-subject emotion. Published studies have already shown that individuals tend to adjust their communication behavior, such as speech and lexical content, by leveraging the mental state and the behavior of one’s conversational partner [47] [48] [49] [50], which is also validated by the experimental results in Section 7.1. This observation further brings us the insight that the adaptation of one’s body gestures to the emotional states of the corresponding conversational partner may occur occasionally, instead of frequently and continuously, in an interaction.

8 Conclusion and Future Work

In this work, we proposed a statistical framework for robustly modeling body gesture dynamics in interpersonal interactions. The proposed framework is composed of three stages. First, we construct a universal background model (UBM) using Gaussian mixture model (GMM) to represent the subject-independent gesture variability. Next, each gesture sequence is described as a concatenation of semantic gesture patterns using a parallel HMM structure. We further fit the segments of each gesture sequence with the UBM, in order to select statistically prominent gesture patterns for the sequence. The dynamics of each gesture sequence are represented by a statistical variation profile computed from the prominent segments, and are further described in a well-defined kernel space. The framework is flexible and general. Each of the components could be individually modified to satisfy the needs of more complex tasks. For example, more advanced techniques, such as deep neural networks [51], could be used for the global gesture variability modeling, when a large amount of data is available.

We evaluated our model in the emotion recognition experiments and compared with three baseline models. We conducted two experiments: intra-subject and inter-subject emotion evaluation. In the experiments, we considered two types of expressive gestures: head motion and hand gesture. According to the experimental results, our proposed UBM-based framework shows superiority over the baseline models in all cases. This could be attributed to the fitting of each gesture sequence to the UBM which unifies the information of different sequences and concentrates essen-
tial dynamics. The statistical fitting may robustly characterize gesture variations by removing individual-specific idiosyncrasies within a sequence. We also observed that the gesture pattern based models generally outperformed the low-level physical gesture dynamics. Though human gestures are highly variable and complex in nature in terms of both spatial and temporal structures, this observation corroborates the flexibility and advantage of representing a complex gesture sequence with elementary gesture patterns. In addition, hand gesture demonstrates a greater ability for expressing emotions compared to head motion. As shown in Fig. 4, head motion is determined by three degrees of freedom, while hand gesture is conditioned on 12 independent variables. The higher degree of hand gesture variability may introduce richer expressiveness during communication. Inspired by this finding, future work on expressive gesture animation could especially focus on synthesizing expressions of hand gesture. Furthermore, it is interesting to observe that our framework also shows effectiveness in the cross-subject emotion evaluation, due to the nature of the complex behavior coordination between the interacting partners.

Analysis of the relationship between the emotion recognition performance and the number of the salient gesture patterns indicates the underlying structure of human summarization process of the global emotion perception. We found that a better recognition performance can be achieved with a few gesture patterns, compared to using the entire gesture sequence. Hence, a few local salient events are sufficiently informative to inform the human summarization of the global emotion perception over an interaction. Another interesting observation is that fewer salient segments are needed to summarize the cross-subject emotion, compared to those used for abstracting the intra-subject emotion. Since cross-subject emotion evaluation implies the behavior adaptation of an individual towards the emotional state of the corresponding conversational partner, this observation brings us the insight that such adaptation may occur sporadically, instead of frequently and continuously, in an interaction.

It is worth noting that the recognition performance in the experiments may be not as high as expected for typical acted interactions. However, unlike other acted data [18] [22], the formal design of the CreativeIT database is based on the theatrical improvisation technique of Active Analysis pioneered by Stanislavsky. The key element of Active Analysis is that actors need to keep a verb (e.g., to persuade or to approach) in mind, which drives their actions during the performance. As a result, different communication manifestations, such as emotions, attitudes, and speech or body gestures, can be naturally elicited through the course of the interaction. The acted interactions in our database are therefore closer to natural interpersonal communication. The elicited naturalness of actors’ behavior induces more difficulty in body gesture modeling, and hence leads to a lower performance than typically expected.

The proposed framework is especially suitable for representing long-term body gestures with high dynamical complexity, such as the interaction-context body gestures in our case. However, it is also applicable to model dynamics of general body gestures that occur in any other contexts. As a future direction, it would be interesting to extend our model particularly for the body gestures from interpersonal communication by incorporating interaction-context factors. For example, we could simultaneously consider the dynamics of the interlocutor while modeling the body gestures of an interaction participant.

Another future direction based on the proposed model could particularly focus on analyzing the role of different body parts in expressing emotions. This direction could be approached by studying the emotion recognition performance with dynamics of joint-body (e.g., head-hand) gesture patterns or with dynamics combinations of different body parts, which could aid in the development of automatic emotion recognition systems as well as human-machine interfaces.

The effectiveness of body gestures for emotion recognition implies the possibility of expressive gesture animation. One long-term goal for future work is to animate body motion for a virtual agent which could express one’s emotional state towards the human interlocutor. Since the proposed framework for gesture dynamics modeling is based on discrete gesture patterns, it could be well incorporated into the traditional pattern-based animation approaches where a complex gesture is generated by continuously concatenating selected patterns. We have also demonstrated that human the global perception of emotions is triggered by a few events over an interaction. Therefore, to achieve a pleasant and natural human-agent conversation, it is also possible to create an intelligent agent which can instantly sense the emotional state of the human user from very few prominent gesture events in the beginning of the conversation, adapt the dialog strategy accordingly and respond appropriately.

References


Zhaojun Yang is a Ph.D. candidate in Electrical Engineering at the University of Southern California (USC). She received her B.E. Degree in Electrical Engineering from University of Science and Technology of China (USTC) 2009 and M.Phil. Degree in Systems Engineering and Engineering Management from Chinese University of Hong Kong (CUHK) 2011. Her research interests include multimodal emotion recognition and analysis, interaction modeling and spoken dialog system. She was awarded with the USC Annenberg Fellowship (2011-2015).
Shrikanth S. Narayanan (S88-M95-SM02-F09)
is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics, Psychology Neuroscience and Pediatrics and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biomedical problems and applications with direct societal relevance. [http://sail.usc.edu]