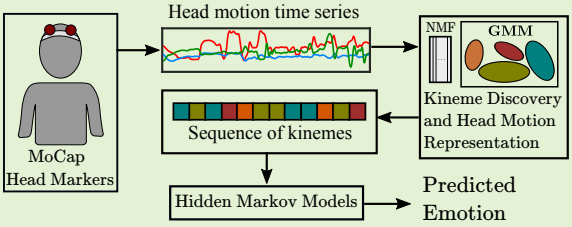
Emotion Sensing From Head Motion Capture

Atanu Samanta, Tanaya Guha, *Member, IEEE*

***Abstract*— Computational analysis of emotion from verbal and non-verbal behavioral cues is critical for human-centric intelligent systems. Among the non-verbal cues, head motion has received relatively less attention, although its importance has been noted in several research. We propose a new approach for emotion recognition using head motion captured using Motion Capture (MoCap). Our approach is motivated by the well known *kinesics-phonetic analogy*, which advocates that, analogous to human speech being composed of phonemes, head motion is composed of *kinemes* i.e., elementary motion units. We discover a set of kinemes from head motion in an unsupervised manner by projecting them onto a learned basis domain and subsequently clustering them. This transforms any head motion to a sequence of kinemes. Next, we learn the temporal latent structures within the kineme sequence pertaining to each emotion. For this purpose, we explore two separate approaches – one using Hidden Markov Model and another using artificial neural network. This class-specific, kineme-based representation of head motion is used to perform emotion recognition on the popular IEMOCAP database. We achieve high recognition accuracy (61.8% for three class) for various emotion recognition tasks using head motion *alone*. This work adds to our understanding of head motion dynamics, and has applications in emotion analysis and head motion animation and synthesis.**



***Index Terms*— Emotion recognition, head motion modeling, hidden Markov model, motion capture, non-negative matrix factorization.**

I. Introduction

COMPUTATIONAL modeling and analysis of emotion from verbal and non-verbal behavioral cues (facial expressions, head motion, body gesture, brain activity) are critical for human-centric systems, such as driver's behavior monitoring [1], social robotics [2] and mental health monitoring [3], [4]. In the last decade, emotion analysis from speech [5] has become a major research topic. Among the non-verbal cues, facial expression has been studied widely for emotion recognition [6]–[8]. Several works have also used body gesture to recognize emotion [9], [10]. Usage of brain activity sensing through electroencephalography (EEG) for emotion recognition is gaining growing attention in recent years [11], [12]. Head motion, however, has received relatively less attention, although its importance has been noted in several research [13]–[16].

A recent study [13] reported that humans can distinguish among sadness, anger and neutral emotion with an accuracy of around 70% by observing the head motion *alone*. Studies have also shown that certain head motion patterns can be effectively associated with particular emotions [17], [18]. Research has also shown that head motion, when compared to facial expressions, contains additional information about human emotions [16], [19]. Despite the evidences about the usefulness of head motion in emotion recognition, efforts to build an effective mathematical model of head motion remain limited.

The majority of the existing works on head motion rely on extracting low-level features from the data. Ding et al. [20] use the amplitude of representative Fourier components to construct a dynamic feature vector. Samanta and Guha [16] proposed to extract energy of displacement, velocity and acceleration of the pitch, yaw and roll of head motion. Gunes and Pantic [21] use low-level features (magnitude and directions of 2D head motion) and higher level features, such as nods and shakes to predict emotion. Yang and Narayanan [22] proposed a statistical distance measure between head motion time series through the extraction of meaningful head gesture segments using parallel HMM and universal background model. Xiao et al. [23] extract optical flow to capture the 2D head motion from video and propose a distance measure between head motion time series. In the context of emotion analysis, head motion has been used to study the coordination between mother and infants [15], [24], interpersonal coordination in couples therapy [4], and to analyse spontaneous affect [25].

In this paper, we propose an unsupervised approach to modeling head motion (captured using motion sensors) in the context of emotion recognition. Our approach is motivated by Birdwhistell's *kinesics-phonetic analogy* [26], which states that just as human speech is composed of *phonemes* (elementary units of verbal language), head motion is composed of short elementary motion units called *kinemes*. The past work by Xiao et al. [23] also follows this analogy. However, they consider only those head motion segments which involve large motion, and discards the rest as non-motion. However, studies

Atanu Samanta is with the Electrical Engineering Department, Indian Institute of Technology, Kanpur, India.

Tanaya Guha is with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

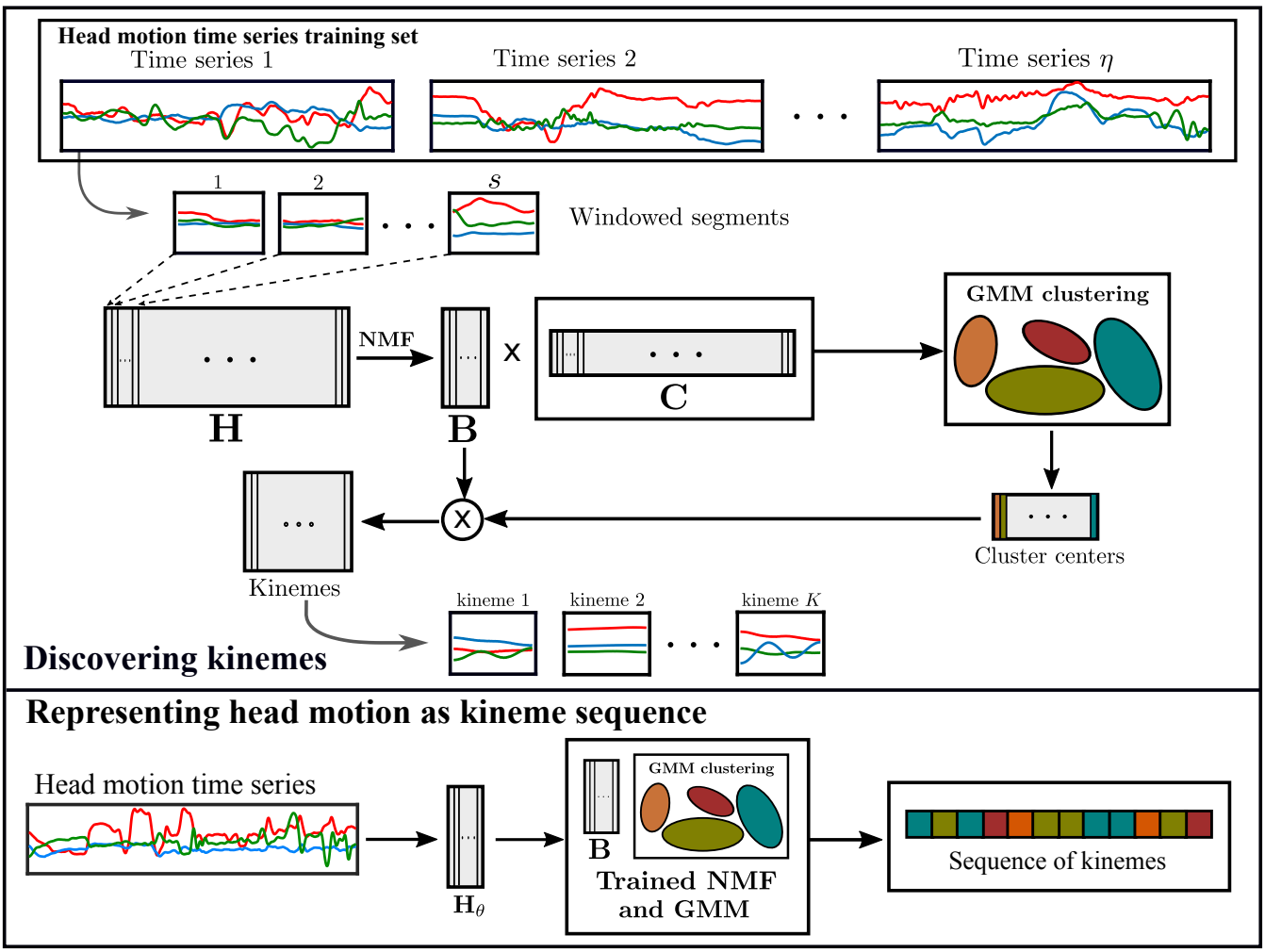


Fig. 1: Overview of the proposed framework of modeling affective head motion.

show that slower and subtle head motion, even still head pose, contain emotion-related information [17], [18]. For example, a leaning forward head movement or slow head movement can be associated with sadness [18], while fear or anger is associated with a head movement of leaning backward [17]. Hence, we consider the entire available head motion, and rely on our unsupervised modeling approach to discover the fundamental motion units. Different from the works of Xiao et al. [23] and Yang and Narayanan [22], we focus on learning the latent temporal structure in kineme sequence while the other [22], [23] simply use distance measures between kinemes or features.

Our model follows a two-stage approach. First, we discover a set of kinemes from head motion data by clustering head motion patterns in a learned basis domain. This transforms head motion data to a sequence of kinemes. Next, we learn the temporal latent structures in the kineme sequence pertaining to each emotion class. For this purpose we explore two different approaches – one using a Hidden Markov Model (HMM) and the other using Long Short Term Memory (LSTM). This class-specific, kineme-based representation of head motion is used to recognize emotion on the IEMOCAP [27] database. Our model achieves an accuracy of 61.8% for recognising emotion (3 classes) using head motion alone. For a similar task, reported human accuracy is $\sim$ 70% [13]. To contextualize how useful head motion as compared to other emotional signals, we compared our results (52.1%) with speech-based recognition results (69.1%). Experiments show that head motion contain significant emotional content, which can improve affective systems.

II. Proposed Framework

In this section, we propose a new framework for modeling head motion motivated by the kinesics-phonetic analogy [26] that advocates representing head motion as a composition of kinemes (analogous to phonemes in verbal languages). We develop a two-stage approach to learn a kineme-based representation of (affective) head motion. Our framework has two major stages. First, in the *kineme discovery* stage, we learn the kinemes (short segments of interpretable motion units) from head motion data in an unsupervised manner. This lets us represent any head motion as a sequence of kinemes. Next, in the *affect-specific kineme representation* stage, we learn the temporal latent structures present in the kineme dynamics for different affective states, and subsequently produce a compact representation of head motion for each affective state. Fig. 1 presents an overview of our proposed approach.

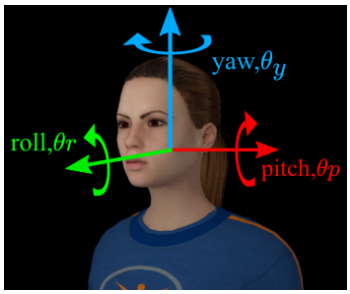


Fig. 2: Head pose defined in terms of the rotation of head about three principal axes – pitch, yaw and roll.

A. Kineme discovery

In this stage, we are concerned with learning a set of fundamental head motion units, referred to as the kinemes. The basic idea is to perform effective clustering of the input head motion segments by projecting them onto a lower dimensional space, so as to yield interpretable and representative head motion patterns.

Head motion representation. Head motion can be seen as the result of continuous rotation of a 3D rigid body i.e., head. For simplicity of representation, we consider only the head rotation in this study and leave out translation. Nevertheless, the framework can easily accommodate head translation, if needed. Head rotation at time instant t is commonly defined in terms of the three Euler angles: pitch θ_p^t, yaw θ_y^t and roll θ_r^t (see Fig. 2). Thus head motion can be defined as a time series of the 3D head poses: $\boldsymbol{\theta} = \left\{\left(\theta_p^t, \theta_y^t, \theta_r^t\right)\right\}_{t=1}^{T}$, where T is the total duration of the head motion time series. The Euler angles are defined in the range between 0° and 360° to ensure non-negativity.

Consider a segment of length L in $\boldsymbol{\theta}$ as $\left[\left(\theta_p^l, \theta_y^l, \theta_r^l\right), \cdots, \left(\theta_p^{l+L}, \theta_y^{l+L}, \theta_r^{l+L}\right)\right]$. This segment can be reordered as $\boldsymbol{\phi} = [\boldsymbol{\phi}_p, \boldsymbol{\phi}_y, \boldsymbol{\phi}_r]$, where $\boldsymbol{\phi}_p = \{\theta_p^t\}_{t=l}^{l+L}$, $\boldsymbol{\phi}_y = \{\theta_y^t\}_{t=l}^{l+L}$ and $\boldsymbol{\phi}_r = \{\theta_r^t\}_{t=l}^{l+L}$. We characterize $\boldsymbol{\phi}$ by a vector $\mathbf{h}$ as follows

$$\mathbf{h} = \left[\boldsymbol{\phi}_p, \boldsymbol{\phi}_y, \boldsymbol{\phi}_r, |\nabla\boldsymbol{\phi}_p|, |\nabla\boldsymbol{\phi}_y|, |\nabla\boldsymbol{\phi}_r|\right]^\intercal, \tag{1}$$

where ∇ denotes first order derivatives. The derivatives are augmented in order to capture additional information about the temporal dynamics of the head motion. Given $\boldsymbol{\theta}$, its characterization matrix $\mathbf{H}_{\boldsymbol{\theta}}$ is defined as

$$\mathbf{H}_{\boldsymbol{\theta}} = [\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \cdots, \mathbf{h}^{(s)}], \tag{2}$$

where $\mathbf{h}^{(i)}$ represents segment $\boldsymbol{\phi}^{(i)}$, and s is the total number of segments obtained using an overlapping window of length L and overlap $L/2$.

Lower dimensional representation learning. Given a training set of η number of head motion time series $\{\boldsymbol{\theta}_j\}_{j=1}^{\eta}$. A head motion matrix containing all the training samples is computed as follows:

$$\mathbf{H} = \left[\mathbf{H}_{\boldsymbol{\theta}_1} | \mathbf{H}_{\boldsymbol{\theta}_2} | \cdots | \mathbf{H}_{\boldsymbol{\theta}_\eta}\right]. \tag{3}$$

Note that each column of $\mathbf{H}$ represents a single segment of a head motion time series. In order to discover kinemes in an unsupervised manner, we propose to first learn a set of independent basis from the training data $\mathbf{H} \in \mathbb{R}_+^{m\times n}$, where $n = \eta s$. Non-negative matrix factorization (NMF) is a popular basis learning approach that is known to generate interpretable basis [28]. As kinemes are considered to be fundamental, interpretable gesture segments, we decompose $\mathbf{H}$ into two non-negative matrices: $\mathbf{B} \in \mathbb{R}_+^{m\times q}$ and $\mathbf{C} \in \mathbb{R}_+^{q\times n}$ as follows.

$$\min_{\mathbf{B}\geq 0, \mathbf{C}\geq 0} \quad \|\mathbf{H} - \mathbf{B}\mathbf{C}\|_F^2, \tag{4}$$

where $q \leq \min(m, n)$ and $\|\cdot\|_F$ denotes the Frobenius norm. The columns of $\mathbf{B}$ are the bases of the q-dimensional subspace and the columns of $\mathbf{C}$ are the q-dimensional representation of the corresponding head gesture segments in $\mathbf{H}$. The optimization problem in (4) is non-convex with respect to $\{\mathbf{B}, \mathbf{C}\}$ but is convex for $\mathbf{B}$ and $\mathbf{C}$ separately. The most popular NMF algorithm to solve (4) involves the multiplicative update rule [29] that updates randomly initialized $\mathbf{B}$ and $\mathbf{C}$ alternatively as follows

$$\mathbf{C} \leftarrow \mathbf{C} \odot \frac{\mathbf{B}^\intercal\mathbf{H}}{\mathbf{B}^\intercal\mathbf{B}\mathbf{C}}, \qquad \mathbf{B} \leftarrow \mathbf{B} \odot \frac{\mathbf{H}\mathbf{C}^\intercal}{\mathbf{B}\mathbf{C}\mathbf{C}^\intercal}, \tag{5}$$

where $\odot$ denotes element-wise product and the division is also an element-wise division. Given that $\mathbf{B}$ is learned from a large training dataset, any new head motion segment $\mathbf{h}_j \in \mathbf{H}$ can be represented using the non-negative bases in $\mathbf{B}$ with acceptable accuracy.

$$\mathbf{h}_j \approx \mathbf{B}\mathbf{c}_j, \tag{6}$$

where $\mathbf{c}_j \in \mathbb{R}_+^p$, the j-th column of the matrix $\mathbf{C} = \{\mathbf{c}_j\}_{j=1}^n$ is the lower dimensional representation of the j-th head motion segment in the matrix $\mathbf{H} = \{\mathbf{h}_j\}_{j=1}^n$.

Clustering to discover the kinemes. Recall that our primary objective is to cluster the head motion segments in $\mathbf{H}$ to discover the kinemes. Instead of clustering the collection of raw segments in $\mathbf{H}$, we cluster their corresponding lower dimensional representations obtained via NMF for higher interpretability and stability.

Given the n head motion segments in $\mathbf{H} = \{\mathbf{h}_j\}_{j=1}^n$, we consider their corresponding lower dimensional representations $\mathbf{C} = \{\mathbf{c}_j\}_{j=1}^n$. To cluster the head motion segments in the space spanned by the NMF bases, we learn a Gaussian mixture model (GMM) $\boldsymbol{\Psi} : \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ such that

$$P(\mathbf{c}_j|\boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{c}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{7}$$

where $\pi_k \in \mathbb{R}_+$, $\boldsymbol{\mu}_k \in \mathbb{R}_+^p$ and $\boldsymbol{\Sigma}_k \in \mathcal{S}_+^p$ are the weight, mean and covariance matrix of the k-th Gaussian, $\sum_{k=1}^K \pi_k = 1$ and K is the number of components (clusters) in the mixture. Our covariance matrices $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ are considered diagonal for computational simplicity. The model parameters $\boldsymbol{\Psi}$ are estimated using standard Expectation Maximization (EM) algorithm, where each component of the mixture model $\boldsymbol{\Psi}$ corresponds to a class of kineme. Each cluster center $\boldsymbol{\mu}_k \in \mathbb{R}_+^p$ best represents the corresponding kineme in the subspace spanned by the NMF bases. However a kineme is understood as a temporal segment in the original head motion subspace characterized by the Euler angles. Hence we compute

the following transformation for each of the K components in the mixture model

$$\tilde{\mathbf{h}}_k = \mathbf{B}\boldsymbol{\mu}_k, \quad k = 1, \cdots K \tag{8}$$

where $\tilde{\mathbf{h}}_k$ is the k-th kineme.

B. Class-specific kineme representation

1) Head motion as a sequence of kinemes: After we learn the kinemes $\{\tilde{\mathbf{h}}_k\}_{k=1}^{K}$, *any* head motion time series $\boldsymbol{\theta}^* = \{\phi^{*(i)}\}_{i=1}^{s}$ can be represented compactly in terms of the K kinemes by associating each of its segment $\phi^{*(i)}$ to an individual kineme. For each $\boldsymbol{\phi}^*$, we compute its corresponding characterization vector $\mathbf{h}^*$ following (1), and then project $\mathbf{h}^*$ onto the learned subspace spanned by $\mathbf{B}$ to obtain $\mathbf{c}^*$ as follows

$$\hat{\mathbf{c}}^* = \arg\min_{\mathbf{c}^* \geq 0} \|\mathbf{h}^* - \mathbf{B}\mathbf{c}^*\|_F^2 . \tag{9}$$

To associate $\hat{\mathbf{c}}^*$ to one of the K kinemes we maximize the posterior probability,

$$\arg\max_{k} \quad P(k|\hat{\mathbf{c}}^*). \tag{10}$$

where $p(k|\hat{\mathbf{c}}^*)$ is given by,

$$P(k|\hat{\mathbf{c}}^*) = \frac{P(k)P(\hat{\mathbf{c}}^*|k)}{P(\hat{\mathbf{c}}^*|\boldsymbol{\Psi})} = \frac{\boldsymbol{\pi}_k \mathcal{N}(\hat{\mathbf{c}}^*; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\hat{\mathbf{c}}^*; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \tag{11}$$

After each $\phi^*(i)$ is mapped to one of the K kinemes, we can represent $\boldsymbol{\theta}^*$ as a sequence of kinemes $\mathcal{K} = [k_1, k_2, \cdots, k_s]$.

2) Kineme representation: Recall our hypothesis that different affective states in head motion can be represented by different temporal patterns in the kineme dynamics. We explore the following two (unsupervised and supervised) models for capturing these latent structures.

Using HMM. We use hidden Markov model (HMM) for capturing the latent structures in the sequence of kinemes by learning a separate HMM model $\{\boldsymbol{\lambda}_e\}_{e=1}^{E}$ for each emotion class $e \in \{1, 2, \cdots, E\}$. Assume that the training set corresponding to e is represented by the set of kineme sequences $\{\mathbf{k}_e^1, \mathbf{k}_e^2, \cdots, \mathbf{k}_e^{N_e}\}$, where N_e is the number of time series pertaining to e. The model parameters $\boldsymbol{\lambda}_e$ are estimated by maximizing the likelihood of the kineme sequences,

$$\boldsymbol{\lambda}_e \quad = \quad \arg\max_{\boldsymbol{\lambda}_e} \quad \prod_{i=1}^{N_e} P(\mathbf{k}_e^i|\boldsymbol{\lambda}_e) \tag{12}$$

We use the popular Baum-Welch iterative algorithm to solve the above maximization problem and train the HMM models for each emotion class.

Using ANN. We employ artificial neural network (ANN) as a second approach to learn the latent structures in the affect specific kineme sequences. Our ANN consists of one hidden layer of *Long Short Term Memory* (LSTM) and an output layer with the number of nodes same as the number of emotions and with 'softmax' activation. We use the categorical cross-entropy as the loss function. This network receives the one-hot encoded kineme sequences as input.

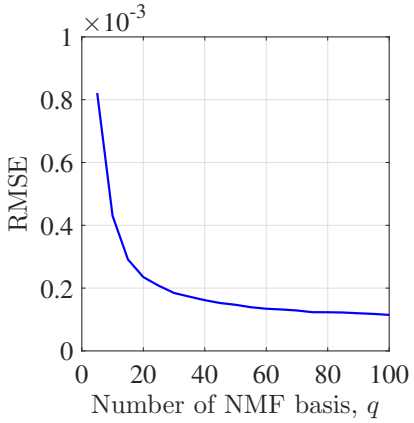


Fig. 3: Reconstruction error vs. number of NMF bases (q).

III. Application to Emotion Recognition

Our proposed model described in the previous section enables us to represent head motion as a sequence of kinemes, and to learn the different kineme dynamics associated with different emotion class. A natural application of this model is in emotion recognition using head motion. To test the strength of our model and to analyze the role of head motion in emotion recognition, we validate our proposed model through a series of emotion recognition experiments on the very popular Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [27].

A. Experimental setup

Given a head motion time series $\boldsymbol{\theta}_t$ our task is to recognize its emotion label. In order to achieve this, we first obtain the windowed segments from $\boldsymbol{\theta}_t$ and create the corresponding characterization matrix $\mathbf{H}_{\boldsymbol{\theta}_t}$ as described in section II-A. Next, we represent $\boldsymbol{\theta}_t$ as a sequence of kinemes $\mathbf{k}_t$ as described in section II-A. Finally we classify the emotions from the kineme sequences using the two models (HMM and ANN) described earlier. We utilize the class-specific HMM models $\boldsymbol{\lambda}_e$ to classify the emotion as the one under which the likelihood of the kineme sequence is maximized,

$$e^* = \arg\max_{e \in \{1,2,\cdots,E\}} P(\mathbf{k}_t|\boldsymbol{\lambda}_e) \tag{13}$$

where, $\{\boldsymbol{\lambda}_e\}_{e=1}^{E}$ are the trained HMM model corresponding to the emotion class $e \in \{1, \cdots, E\}$.

Database. The popular IEMOCAP database [27] provides audiovisual recordings of both interlocutors and the Motion Capture (MoCap) data of the face, head and hand of one of the interlocutor in dyadic interactions. Each utterance of the interactions are annotated into categorical (*neutral, joy, sadness, anger* and *others*) and dimensional emotion labels (*valence, arousal* and *dominance* rated in a scale between 0 and 6). We perform various classification experiments using both the categorical and dimensional labels provided in the database. For simplicity and interpretability, the dimensional space is clustered to generate discrete class labels. For the experiments using categorical labels, we simply retain the data with the four basic emotion labels mentioned above and discard others as for those very few head motion time series

TABLE I: Emotion recognition results on the IEMOCAP database with dimensional labels. Our kineme-based model outperforms both the baselines in all classification tasks.

Method	P	R	Acc↑	Avg F1↑	κ ↑
+ve vs -ve valence					
HMM baseline					
+ valence	0.40	0.38	58.5%	0.54	0.08
- valence	0.68	0.70			
SVM baseline					
+ valence	0.40	0.40	57.8%	0.54	0.08
- valence	0.68	0.68			
Kineme + HMM (ours)					
+ valence	0.45	0.62	60.2%	0.59	0.20
- valence	0.74	**0.59**			
Kineme + LSTM (ours)					
+ valence	**0.46**	**0.66**	**60.5%**	**0.60**	**0.21**
- valence	**0.76**	0.58			
High vs low arousal					
HMM baseline					
High arousal	0.64	0.36	47.8%	0.48	0.03
Low arousal	0.39	0.68			
SVM baseline					
High arousal	0.64	0.49	51.5%	0.51	0.04
Low arousal	0.40	0.56			
Kineme + HMM (ours)					
High arousal	0.75	0.83	72.9%	0.70	0.41
Low arousal	0.67	**0.56**			
Kineme + LSTM (ours)					
High arousal	**0.76**	**0.90**	**76.2%**	**0.73**	**0.47**
Low arousal	**0.77**	0.54			

P = Precision, **R** = Recall, **Acc** = Accuracy, κ = Cohen's kappa score

TABLE II: Emotion recognition results on the IEMOCAP database with categorical labels. Our kineme-based model outperforms both the baselines in all classification tasks.

Method	P	R	Acc↑	Avg F1↑	κ ↑
Three categorical emotions					
HMM baseline					
Sadness	0.64	0.34			
Anger	0.43	0.42	36.2%	0.35	0.06
Neutral	0.12	0.35			
SVM baseline					
Sadness	0.69	0.77			
Anger	0.24	0.17	53.9%	0.37	0.11
Neutral	0.17	0.17			
Kineme + HMM (ours)					
Sadness	**0.85**	0.59			
Anger	**0.45**	**0.81**	59.8%	**0.53**	0.35
Neutral	0.30	**0.30**			
Kineme + LSTM (ours)					
Sadness	**0.85**	**0.65**			
Anger	0.43	0.78	**61.8%**	**0.53**	**0.36**
Neutral	**0.38**	0.26			
Four categorical emotions					
HMM baseline					
Sadness	0.54	0.28			
Anger	0.47	0.22	30.2%	0.29	0.07
Neutral	0.12	0.26			
Joy	0.20	0.65			
SVM baseline					
Sadness	0.63	0.70			
Anger	0.33	0.22	47.9%	0.33	0.14
Neutral	0.17	0.17			
Joy	0.22	0.24			
Kineme + HMM (ours)					
Sadness	0.81	**0.59**			
Anger	**0.40**	0.64	**52.1%**	**0.41**	**0.29**
Neutral	0.33	**0.30**			
Joy	0.13	**0.18**			
Kineme + LSTM (ours)					
Sadness	**0.87**	0.51			
Anger	0.34	**0.83**	50.3%	0.39	0.28
Neutral	**0.35**	0.26			
Joy	**0.18**	0.12			

P = Precision, **R** = Recall, **Acc** = Accuracy, κ = Cohen's kappa score

could be generated.

Data preprocessing. For our experiments, we only use the MoCap data (for head motion, recorded at 120 frames/second) of a subject if the emotion label is provided for the utterance. Note that the head motion time series, owing to the different lengths of the utterances, are originally of different lengths. We create a set of 299 time series of uniform length (10 seconds) by subdividing those longer than $T = 10$ seconds into multiple segments, and by discarding those shorter than 10 seconds. Thereafter all head motion time series are further segmented using an window size $L = 1$ second with 50% overlap as described in section II-A. The parameters $T = 10$ seconds and $L = 1$ second are chosen experimentally. The effect of these parameters on the overall emotion recognition performance are discussed later in Section III-C.

Evaluation metrics. Our database has high class imbalance. Therefore we use precision, recall, F1-score and kappa score (κ) alongside accuracy for evaluation in a 10-fold cross-validation setting.

Baselines. In order to validate the effectiveness of the proposed kineme-based model, we create two simple baselines that performs classification directly on the raw head motion data. **(a) Support vector machine (SVM) baseline:** We take each 10 second long head motion segment $\boldsymbol{\theta}$ as a data point and represent it using its corresponding characterization vector $\mathbf{h} = [\boldsymbol{\theta}_p, \boldsymbol{\theta}_y, \boldsymbol{\theta}_r, |\nabla\boldsymbol{\theta}_p|, |\nabla\boldsymbol{\theta}_y|, |\nabla\boldsymbol{\theta}_r|]^\intercal$. This data is input to an SVM classifier to perform emotion recognition under the same cross-validation setting as above. **(b) HMM baseline:** We learn HMM models for each class using raw head motion time-series (as opposed to kineme representation as in our proposed method) as training data. The label of a test data is given by the HMM model with highest likelihood.

Choosing the number of NMF bases. The number of NMF bases is set to $q = 20$ in our experiments. This number of course is a user defined parameter. Fig. 3 shows the reconstruction error incurred on the data matrix $\mathbf{H} \in \mathbb{R}_+^{m \times n}$ while varying the number of bases q. The reconstruction error is measured in terms of root mean square error (RMSE) as $\frac{1}{mn} \|\mathbf{H} - \mathbf{BC}\|_F$, where $\mathbf{B} \in \mathbb{R}_+^{m \times q}$ and $\mathbf{C} \in \mathbb{R}_+^{q \times n}$ are the factor matrices. Fig. 3 shows that $q = 20$ is the knee point, i.e., NMF with $q = 20$ offers the best trade off between the reconstruction error and the number of bases.

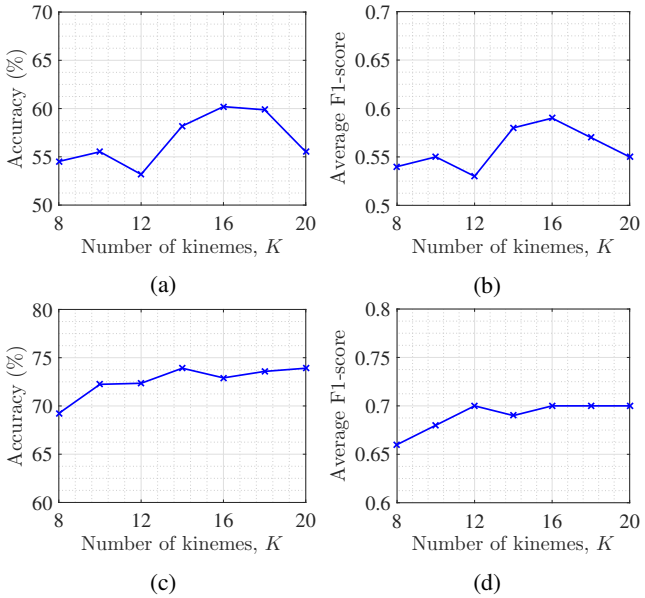


Fig. 4: Dependency of emotion classification performance (accuracy and F1-Score) on the number of kinemes K. (a)-(b) *valence classification*, (c)-(d) *arousal classification*.

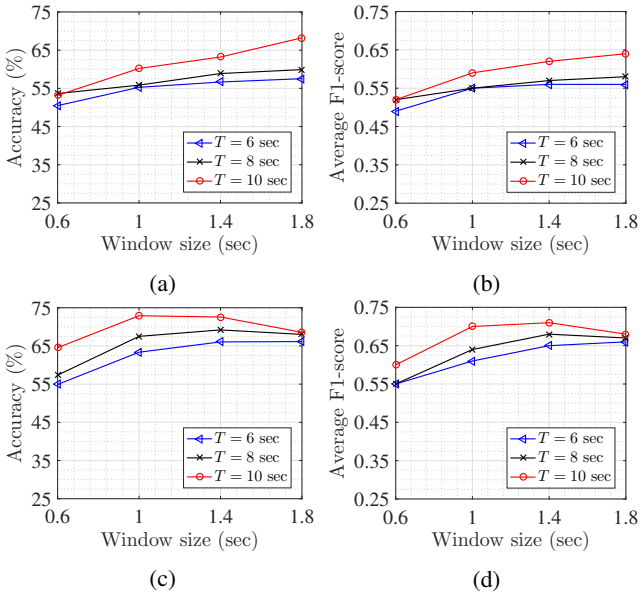


Fig. 5: Dependency of emotion classification performance (accuracy and F1-Score) on the head motion time series length and window size. (a)-(b) *valence classification*, (c)-(d) *arousal classification*.

B. Performance evaluation

We perform various classification experiments using both the categorical and dimensional labels provided in the IEMOCAP database. Since little is known about the role of head motion alone in classifying emotions, we perform various emotion recognition tasks by systematically varying the difficulty level of the tasks. For all our experiments, the number of kinemes (i.e., the number of GMM components) is set to $K = 16$. The effect of varying the number of kinemes K on the overall performance is discussed later in section III-C.

Experiments using dimensional labels. Our first task is to perform binary classification to distinguish between (i) positive (105 samples) and negative valence (194 samples) and (ii) high (185 samples) and low arousal (114 samples). We partitioned all the samples in the database into two groups directly using the valence or arousal ground truth labels. For each dimension, the grouping was done by choosing a threshold that is the mean of the maximum and the minimum annotation range used by the annotators (not necessarily equal to the min/max value possible). The results are shown in Table I. The order of the HMM is set separately for each emotion class. The class with more complexity and variability requires higher order HMM. Our kineme-based model outperforms both the baselines in all three classification tasks in terms of the accuracy, the average F1-score and the kappa score (κ). Notice that all the models (including our proposed model) perform well in classifying between high and low arousal. High arousal is often accompanied with larger head motion, which has been efficiently captured by the models to distinguish from lower arousal behavior. Valence has been classified with the lowest accuracy (and F1-score and κ) among other dimensions. This is aligned with observations made in past works regarding lower recognition accuracy of valence in facial expressions and speech [30].

TABLE III: Comparison with other modalities on the IEMOCAP database for 4 class emotion recognition accuracy (%).

Modality	Sadness	Anger	Neutral	Joy	All
Speech [31]	73.3	80.2	65.9	37.5	69.1
Speech + Face [32]	77.8	77.2	46.9	68.4	67.6
Head motion (ours)	59.1	63.9	30.4	17.6	52.1

Experiments using categorical labels. Next we perform two classification experiments using the categorical labels. For these experiments, we have considered the four basic emotion categories: *anger, happiness, neutral* and *sadness* with 93, 36, 23 and 17 samples respectively. Since the *happiness* class has very few (only 17 time series) data compared to the other three, we report classification results with and without the happiness class in Table II. Our proposed kineme-based model outperforms both the baselines for both the tasks in terms of accuracy, average F1 score and κ. Our kineme-based model significantly outperforms the HMM baseline in all experiments (both categorical and dimensional), which establishes the effectiveness of the proposed kineme representation over using raw head motion.

For the 4-class classification task, our model achieved a recognition accuracy of 52.1%. We compare this result with those achieved using other modalities in a similar set up. The comparison in Table III shows that head motion alone contains significant information to detect emotion, and performs well when put in the context of more expressive modalities such as speech and facial expressions. Table IV places our results alongside the reported emotion recognition accuracy of related work on head motion and the human accuracy. Note that this should not be used for one-to-one comparison since

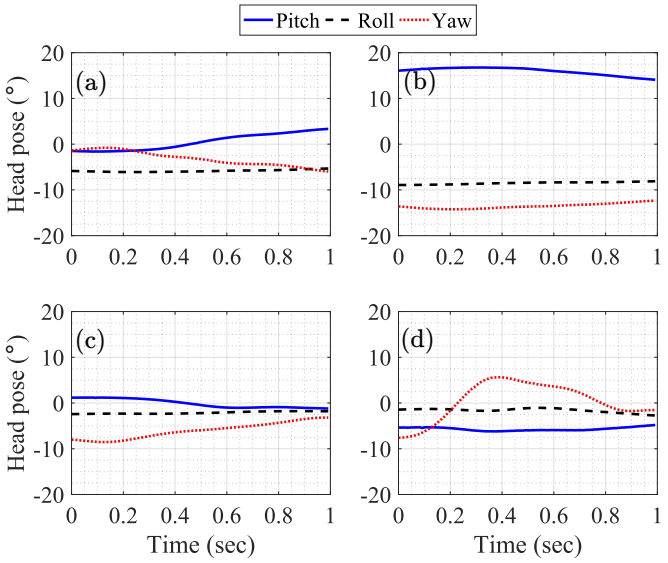


Fig. 6: Temporal waveform of the 4 selected kinemes out of 16 kinemes discovered by our model from the IEMOCAP dataset.

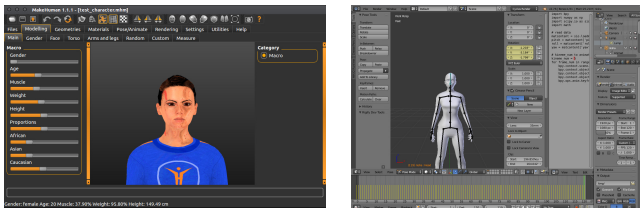
these numbers were reported for different tasks on different databases.

The kineme + LSTM approach outperforms the kineme + HMM approach in every classification tasks except for classifying 4 categorical emotions. It is worth observing that in this classification task one emotion (happiness) have a very few (only 17) time series which makes it difficult for the LSTM to learn any meaningful pattern.

C. Analyzing the effects of parameters

The number of kinemes (K). The number of kinemes is equivalent to the number of components in the GMM model. This is a critical parameter for our model. To study the effect of this parameter on recognition accuracy, we conduct all of the above classification experiments with varying values of K. Fig. 4 shows how the recognition accuracy and average F1-score vary with number of kinemes. Fig. 4 indicates that the performance of our model is relatively more sensitive to the choice of K for valence classification. Overall, it is clear $K = 16$ is a suitable value, implying that only 16 basic kinemes are sufficient to model the affective head motion. Due to space constraint, we present results for only two classification tasks, but similar trends are observed for the others.

We thus learn 16 kinemes (the number of GMM clusters) from the data as described in the section II-A for all experiments. Fig. 6 shows the representative head motion time series pertaining to the 16 kinemes discovered from the training data. Since the NMF decomposition is not unique the learned kinemes are also not unique. In every simulation the discovered set of kinemes may be slightly different. Nevertheless, a close observation of Fig. 6 and the animated videos provided as the supplementary material reveals that the discovered kinemes closely resemble some of the typical head gesture patterns, such as nods, and sweeping head from left to right. We also note that few kinemes represent very slow movement of head - almost a still head pose. This



(a) (b)

Fig. 7: Screenshot from our animated kineme videos in making: (a) creating 3D human using Makehuman, and (b) adding the motion learned from our model using Blender

TABLE IV: Reported emotion recognition accuracy of various methods using only head motion. *This data should not be used for comparison since these were reported for different tasks for different database. Purpose of this table is only to contextualize our results.*

Method	Task	Accuracy
Human observation [13]	classification of 3 categorical emotions	68.5%
Xiao et al. [23]	binary classification of the presence of various behavioral code	57% - 64%
Yang and Narayanan [22]	3 classes (clustered VA space)	61.0%
	4 classes (clustered VA space)	50.5%
Our method	3 categorical emotions	61.8%
	4 categorical emotions	52.1%

can be explained by the fact that our dataset contains many samples of related to sadness, and sadness is known to be associated with slow head movement [18].

Time series length (T) and Window size (L). In the proposed approach, a head motion time series of length T is segmented using an overlapped window of length L, which in turn determines the length of kinemes. These two parameters together determine the scale at which kinemes are discovered. In order to study the effect of these two parameters on the overall performance, we experimented with a range of window size L (0.6 to 1.8 second) for three different values of T (6, 8 and 10 second). Fig. 5 presents the corresponding accuracy and the average F1-score of the emotion recognition tasks. Fig. 5 indicates that larger T values improve classification accuracy. This simply explains that a longer head motion time series provides more segments to better learn the kinemes, which improves classification accuracy. Fig. 5 also shows that for $L = 1$ and $L = 1.4$ the accuracy is higher. This imply that this length is suitable to model kinemes for emotion analysis.

D. Visualizing the kinemes

Fig. 6 shows the waveforms of the 4 selected kinemes out of 16 kinemes (elementary head motion unit) learned by our model. Here, we visualize those kinemes as animated head motion video clips for interpretability and understanding. For this purpose, we have used two open source softwares – Makehuman [33] and Blender [34]. We have created the human character using Makehuman and then animated the head motion of the Makehuman character using Blender (see Fig 7 for screenshots). We animated the 4 selected kinemes

and the full videos are attached as supplementary files. As observed from the animations, each kineme corresponds to an intepretable head movement, such as a head swing from right to left or from up to down.

IV. Conclusion

We proposed a framework for modeling head motion with application to emotion recognition. Our framework is motivated by Birdwhistell's kinesics-phonetic analogy which advocates representing head motion as a composition of kinemes (fundamental head motion units). We proposed an unsupervised approach to automatically discover the kinemes so as to represent any head motion as sequence of kinemes. However, the kinemes our method discovers are not unique. Note that unlike phonemes the set of kinemes are unknown. Thus it is not possible to directly assess if the discovered kinemes are the 'real' kinemes. Hence we evaluated the model by visualizing (animating) the kinemes, and by its ability to represent head motion accurately for tasks such as emotion recognition.

We employed the proposed model to perform emotion recognition on a benchmark database. Results showed that our kineme-based modeling approach can recognize emotion with high accuracy using head motion information alone. Our method achieved 61.8% accuracy in classifying 3 categorical emotions while the reported human accuracy for a similar task is $\sim 70\%$. Our method is simple and it offers human interpretable features (i.e. elementary head motion units or kinemes). However, the limitation of our method is that all the kinemes it discovers are of the same length. On the other hand our method of representing head motion as kineme sequence is generative at every stage to find application in affect specific realistic head motion synthesis, possibly conditioned on speech. This will be pursued as a future work.

References

[1] A. Mann and P. Ellegaard, "System and method for driver risk assessment through continuous performance monitoring," Sep. 14 2017, uS Patent App. 15/455,058.

[2] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for childrens second language skills," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[3] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, 2018.

[4] Z. Hammal, J. F. Cohn, and D. T. George, "Interpersonal coordination of head motion in distressed couples," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 155–167, 2014.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[6] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[7] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 10, pp. 1683–1699, 2007.

[8] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 1, pp. 28–43, 2012.

[9] F. Noroozi, D. Kamińska, C. Corneanu, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[10] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 40, no. 12, pp. 3007–3021, 2018.

[11] P. Lakhan, N. Banluesombatkul, V. Changniam, R. Dhithijaiyratn, P. Leelaarporn, E. Boonchieng, S. Hompoonsup, and T. Wilaiprasitporn, "Consumer grade brain sensing for emotion recognition," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9896–9907, 2019.

[12] P. Sawangjai, S. Hompoonsup, P. Leelaarporn, S. Kongwudhikunakorn, and T. Wilaiprasitporn, "Consumer grade eeg measuring sensors as research tools: A review," *IEEE Sensors Journal*, vol. 20, no. 8, pp. 3996–4024, 2020.

[13] S. R. Livingstone and C. Palmer, "Head movements encode emotions during speech and song." *Emotion*, vol. 16, no. 3, p. 365, 2016.

[14] P. Ekman, "Differential communication of affect by head and body cues." *Journal of personality and social psychology*, vol. 2, no. 5, p. 726, 1965.

[15] Z. Hammal, J. F. Cohn, C. Heike, and M. L. Speltz, "What can head and facial movements convey about positive and negative affect?" in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 281–287.

[16] A. Samanta and T. Guha, "On the role of head motion in affective expression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2886–2890.

[17] M. Lhommet and S. C. Marsella, "Expressing emotion through posture," *The Oxford Handbook of Affective Computing*, p. 273, 2015.

[18] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Methodology for assessing bodily expression of emotion," *Journal of Nonverbal Behavior*, vol. 34, no. 4, pp. 223–248, Dec 2010. [Online]. Available: https://doi.org/10.1007/s10919-010-0094-x

[19] A. Adams, M. Mahmoud, T. Baltrušaitis, and P. Robinson, "Decoupling facial expressions and head motions in complex emotions," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 274–280.

[20] Y. Ding, L. Shi, and Z. Deng, "Low-level characterization of expressive head motion through frequency domain analysis," *IEEE Transactions on Affective Computing*, 2018.

[21] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *International conference on intelligent virtual agents*. Springer, 2010, pp. 371–377.

[22] Z. Yang and S. S. Narayanan, "Modeling dynamics of expressive body gestures in dyadic interactions," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 369–381, 2017.

[23] B. Xiao, P. Georgiou, B. Baucom, and S. S. Narayanan, "Head motion modeling for human behavior analysis in dyadic interaction," *IEEE transactions on multimedia*, vol. 17, no. 7, pp. 1107–1119, 2015.

[24] Z. Hammal, J. F. Cohn, and D. S. Messinger, "Head movement dynamics during play and perturbed mother-infant interaction," *IEEE transactions on affective computing*, vol. 6, no. 4, pp. 361–370, 2015.

[25] P. Liu and L. Yin, "Spontaneous facial expression analysis based on temperature changes and head motions," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–6.

[26] R. Birdwhistell, "Kinesics and context, essays on body-motion communication," *Philadel-phia: University of Pennsylvania Press*, 1970.

[27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[28] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[29] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[30] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 33–40.

[31] K. Mangalam and T. Guha, "Learning spontaneity to improve emotion recognition in speech," in *Proc. Interspeech 2018*, 2018, pp. 946–950. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1872

[32] Y. Kim and E. M. Provost, "Leveraging inter-rater agreement for audio-visual emotion recognition," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 553–559.

[33] *Makehuman*. [Online]. Available: www.makehumancommunity.org

[34] *Blender*. [Online]. Available: www.blender.org