HHS Public Access

Author manuscript

IEEE Trans Multimedia. Author manuscript; available in PMC 2016 July 13.

Published in final edited form as:

IEEE Trans Multimedia. 2015 July 13; 17(7): 1107–1119. doi:10.1109/TMM.2015.2432671.

Head Motion Modeling for Human Behavior Analysis in Dyadic Interaction

Bo Xiao [Student Member, IEEE],
Signal and Image Processing Institute, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089 USA.

Panayiotis Georgiou [Senior Member, IEEE],
Signal and Image Processing Institute, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089 USA.

Brian Baucom, and
Department of Psychology, University of Utah, Salt Lack City, UT, 84112 USA.

Shrikanth S. Narayanan [Fellow, IEEE]
Signal and Image Processing Institute, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089 USA.

Abstract

This paper presents a computational study of head motion in human interaction, notably of its role in conveying interlocutors' behavioral characteristics. Head motion is physically complex and carries rich information; current modeling approaches based on visual signals, however, are still limited in their ability to adequately capture these important properties. Guided by the methodology of kinesics, we propose a data driven approach to identify typical head motion patterns. The approach follows the steps of first segmenting motion events, then parametrically representing the motion by linear predictive features, and finally generalizing the motion types using Gaussian mixture models. The proposed approach is experimentally validated using video recordings of communication sessions from real couples involved in a couples therapy study. In particular we use the head motion model to classify binarized expert judgments of the interactants' specific behavioral characteristics where entrainment in head motion is hypothesized to play a role: *Acceptance*, *Blame*, *Positive*, and *Negative* behavior. We achieve accuracies in the range of 60% to 70% for the various experimental settings and conditions. In addition, we describe a measure of motion similarity between the interaction partners based on the proposed model. We show that the relative change of head motion similarity during the interaction significantly correlates with the expert judgments of the interactants' behavioral characteristics. These findings demonstrate the effectiveness of the proposed head motion model, and underscore the promise of analyzing human behavioral characteristics through signal processing methods.

B. Xiao, P. Georgiou, and S.S. Narayanan. Address: University of Southern California, 3740 McClintock Avenue, EEB 430, Los Angeles, CA 90089. boxiao@usc.edu, georgiou@sipi.usc.edu, shri@sipi.usc.edu. Phone: 213-740-6432. Fax: 213-740-4651.
B. Baucom. Address: University of Utah, 380 S. 1530 E., BEHS 502, Salt Lake City, UT 84112. brian.baucom@psych.utah.edu. Phone: 801-581-7109. Fax: 801-581-5841

Index Terms

Head motion; Behavioral characteristics; Entrainment; Kinesics; Gaussian mixture model; Linear predictive analysis

I. Introduction

Head motion is an important part of nonverbal communication in human interaction. There have been several classifications of head movement such as based on action type including nodding, shaking, tilting, tossing, dipping, thrusting, dropping, *etc*. [1]; based on frequency, amplitude, continuity and other factors [2]; based on timing, stress, juncture and disfluencies in speech, as well as the meaning or intension while listening [3], [4], [5], [6]. Additionally head motion has been studied in relation to semantics, discourse, and communicative functions [7].

Given the importance of head motion as a communicative and social interaction cue, it is also very important in human behavior analysis. However, due to the seemingly unstructured nature of head motion, it is difficult to quantify behaviors from this modality. A well known coding scheme due to Ekman [8] focuses on function rather than movement characterization. Birdwhistell [9] on the other hand, focuses on characterizing the structural-compositional aspects of the movement, akin to the phonemes (elements of language's phonology such as vowels and consonants) of language. This "kinesic-phonetic analogy" hypothesizes elementary motion units called "kinemes". The drawback of Birdwhistell's scheme is that it requires a meaningful discretization of the kinetic space; unlike natural spoken language that is governed by the rules of fairly well understood grammar, body and head movements are less structured, and do not lend themselves easily to unique and meaningful quantizations.

Although many successful approaches have been reported, the current computational approaches for modeling head motion are still not adequate in meeting the sophisticated needs of psychological research, nor are they adequate in capturing the complex details of head motion and the richer information conveyed therein. A topic that requires further research has been the categorization of head motion. People usually only consider nodding and shaking but have largely neglected others [10], including ignoring attributes such as the magnitude and speed of head motion. In addition, head motion behavior has been less studied in real interpersonal interaction scenarios. Finally, the link between head motion and interactants' behavioral characteristics has not been widely analyzed.

The main contributions of this work include, first, the proposal of a categorical head motion representation obtained in a data driven way; second, using the head motion model as a middle layer construct to link low level head motion signals with high level, *summative* assessment of relevant target behavioral characteristics; and third, analysis of the relation between dyadic head motion entrainment and global behavioral characteristics using the proposed categorical representation framework. Note that in many real applications including the one in this work, only a single overall assessment is provided for an entire relatively long interaction, without direct short-term low level annotations. In such cases it

becomes challenging to directly find the relation between very detailed observational signals and subjective global assessments. Therefore, we aim to create a middle layer of motion patterns that has veritable relations to both observed motion signals and high-level behavioral annotations.

In this paper, we first review related background work — both conceptual and computational in Sec. II. We then propose the head motion model in Sec. III. Specifically, we begin by detecting head movement and computing the optical flow of head motion. We use the Line Spectral Frequencies (LSFs) of the optical flow signals as features that represent the properties of head motion. The key idea is to cluster head motion in an unsupervised way, and we use Gaussian Mixture Model (GMM) of LSF features to provide a generative probabilistic interpretation of head motion events. Ideally, each mixture component would correspond to a kineme realization, and the mixture components can be learned from large amounts of data.

In addition, based on the head motion model, we describe an algorithm to measure behavioral similarity in Sec. IV. Behavioral entrainment [11] is an underlying mechanism in human interactions that relates to affect and clinical outcome particularly in psycho-therapy [12], [13], [14], the domain of interest in the present work. We approximate behavioral entrainment with measures of similarity of the signals that the interlocutors generate. We define a head-motion similarity measure for a pair of motion events using Kullback-Leibler divergence. We estimate the similarity of two time periods by averaging a more behaviorally meaningful part of the pair-wise divergences.

We proceed with a description of the couple therapy corpus that is used for evaluation of our model in Sec. V. We apply the proposed model in two behavior modeling case studies in order to investigate its effectiveness in Sec. VI. We first use the proposed model to predict expert-annotated codes of interlocutors' behavioral characteristics. We then test the correlation of the head motion similarity measure and the behavior codes, as the correlation between behavioral similarity and interlocutors' relationship is of interest to domain experts [14], [15]. Since it is difficult for human to label entrainment, such correlation serves as an indirect way to investigate entrainment.

We discuss our findings and remarks on future work in Sec. VII, and offer our conclusions in Sec. VIII.

This work contributes significant new material over our past work [16], [17]. We employ more than twice of the sessions from the interaction corpus compared to past work; we use *disjoint* sets of data to learn the head motion model and to evaluate its relation to behavioral characteristics; we examine the behavior code classification problem by automatically selecting the parameters for the head motion model; we use symmetric divergence between the interactants for head motion similarity measure, which is easier to interpret; and we analyze the properties of the obtained head motion clusters to show how they conceptually relate to domain knowledge, which further supports the modeling approach.

II. Related work

A. Communicative aspects of head motion

Head motion is an integral part of nonverbal human interaction. Nodding and shaking, referred to as “emblems” [8], are the most typical forms of head movement, although there are many others such as tilting, tossing, dipping, thrusting, dropping, *etc*. [1]. They also vary in terms of frequency, amplitude, continuity and other factors. Hadar *et al*. [2] studied head motion in natural human interactions recorded by polarised-light goniometer that measured the angle of head turning. Based on their results, they suggested five classes of head motion. The frequency of head motion clustered into three classes — slow, ordinary and rapid. In addition, large amplitude linear movement was named posture shift, while small amplitude quick movement was named tremor. Following this work, they also found the relation of head motion patterns to timing, stress, juncture and disfluencies in speech, as well as the meaning or intention while listening [3], [4], [5], [6].

B. Head motion modeling and Behavioral Signal Processing

From an engineering perspective, there have been many studies that involve analyzing, modeling and synthesizing human head behavior. One topic is the estimation of head pose through computer vision techniques [18]. The goal is to infer the orientation of the head either in discrete intervals, or continuously in 3D space, from digital images. For a video sequence, one could track the pose along time [19]. In addition, head tracking has been addressed by methods such as Kalman filtering and particle filtering [20], [21], [22], [23]. Recently, the availability of easy to use depth sensors such as Kinect has helped head pose estimation and tracking [24], [25]. Accurate pose estimation and tracking are very useful for analyzing head motion; however, it does not provide direct information regarding what type of motion is involved.

Other researchers have investigated head motion in order to discern specific patterns, usually “nodding” *vs*. “shaking”. An early study used infrared LEDs to track eye pupils and then used it to detect head nods and shakes with a Hidden Markov Model (HMM) [26]. Nodding and shaking detection using HMMs have also been applied to video data, enabled by face and eye detection [27], [28]. In recent years, multimodal methods of head motion detection have been proposed that take context into consideration, *e.g.*, speaking state [29], and lexical and prosodic features [30]. The analysis of head motion also relates to the modeling of user attitude and emotional states [10], [31], [32].

Existing literature exhibits a clear trend in the multifaceted nature of studies on head motion: from a single visual modality to multimodal data use, from individual behavior to interaction, from modeling the signal itself to the signal’s implication of users’ mental, emotional, and rapport states. The emerging field of study — “Behavioral Signal Processing” (BSP) — encompasses these trends in the study of human behavior within which the present paper is situated. BSP refers to “techniques and computational methods that support the measurement, analysis, and modeling of human behavior signals that are manifested in both overt and covert multimodal cues (expressions), and that are processed and used by humans explicitly or implicitly (judgments and experiences)” [33]. The

essential goal of BSP is to inform human assessment and decision making, providing a computational ancillary for human behavior analysis in real world applications, such as marital psycho-therapy [34], addiction intervention [35], [36], and autism diagnosis [37], [38]. A schematic of the BSP framework adopted in the present paper is shown in Fig. 1.

A notable research topic closely related to BSP is Social Signal Processing (SSP), which focuses on modeling, analysis and synthesis of human social behavior through multimodal signal processing [39]. Theories and methods developed for BSP or SSP are largely shareable, since they both model human behavioral cues.

C. Psychological studies of kinesics

In the psychology field, one of the well known coding systems for human action was proposed by Ekman and Friesen [8]. It includes five classes: *emblems*, *illustrators*, *regulators*, *adaptors* and *affective displays*. The system focused more on the property and interactive function of action rather than directly characterizing the movement.

Birdwhistell suggested an structural and descriptive approach [9], treating nonverbal behavior just like verbal language, with the intuition of a “kinesic-phonetic analogy”. In his definition, a most elementary unit of motion was described as a *kineme*, like a phoneme (element of language’s phonology such as a vowel or a consonant), which were combined to form *kine-morphs*, or even larger units of *kinemorphic constructions*. For example, a head sweep to the left could be one kineme. In this way, Birdwhistell’s hierarchical compositional system of movement is just like that of verbal language.

Unfortunately, such a system has not been widely used in practice due to the difficulty of applying the coding on real data. Unlike speech, nonverbal language is less structured, and hence difficult to precisely discretize into kinemes. It also hence poses difficulty for human coders to agree on the coding assignment. As Harrigan [1] pointed out, the coding systems of head movement were “varied, rarely well-defined, and, with few exceptions, [are] not often organized conceptually or theoretically”.

Can computational technologies help at this point? As Kendon commented [40], Birdwhistell was probably ahead of his time. Advances in signal processing and computer vision techniques offer new opportunities to examine his theory in practice. This motivates us to revisit the Kinesics theory that Birdwhistell proposed. Such an approach is appealing to engineers for its natural connection with visual signals, and the hierarchical and compositional structure that is conducive to model construction. While human ratings can differ due to subjective opinions, annotator background, reliability of the annotator due to fatigue or training, computers remain consistent after training (albeit error prone based on training quality and coverage). Furthermore, computers can in certain cases access a much larger amount of data to obtain a more comprehensive model; for instance through cross-referencing data seen in other sessions remote in time and space.

D. Structure models in computation

Discrimination of human action has been studied extensively in the computer vision domain [41] for applications such as video analysis, retrieval, surveillance, and human-computer

interaction. However, these do not address the problem of finding gesture types since the models only discern actions in pre-defined sets. An unsupervised system that recognizes human actions has been proposed based on Latent Dirichlet Allocation (LDA) model [42]. However, it was only applied to distinguish articulated bodily actions while head motion is much more subtle.

The Hierarchical Dirichlet Process — Hidden Markov Model (HDP-HMM) [43] is appealing to decode head motion. An extended form of the model named HDP-AR-HMM has been shown effective to the problem of speech diarization [44]. However, the results of our preliminary experiments were not encouraging, since we observed that usually the model generated one dominating cluster of motion, and very small remaining clusters. Such degraded performance might be due to difficulty of finding good parameters, prior type and sampling method, or it could be a result of inherent less discriminative nature of head motion. Although typical movements like nodding and shaking were well distinguished, in real interaction scenarios many movements did not fall into prototypical motion categories. Therefore one cluster tended to capture a disproportionate amount of movements.

Sargin *et al.* have proposed a method modeling head motion with Parallel-HMM [45]. The graphical structure was made up by several parallel left-right HMMs that shared common start and end nodes, where the end loops back to the start. The idea was that each branch of the graph captured one type of head motion. However, the method was initially proposed as dependent on the subject, as it was found that the model changed significantly if it was trained on a different subject.

In sum, a variety of methods for modeling head (body) motion exist in literature, yet the problem of establishing a generic structural model of head motion is still challenging.

E. Multimodal behavior analysis

The specific application domain considered in this paper is behavioral coding of interactions related to marital therapy. Even within this domain, there are a number of distinct types of behavioral codes used by the experts to describe different behaviors of interests related to a range of research and clinical questions (See Sec. V-A for more details). Each code description is cued by diverse, and often multiple, relevant communication and interaction modalities: voice, language use, nonverbal vocal and visual cues, *etc.* For example, past work on inferring “high” *vs.* “low” score on behavior codes related to *Blame* patterns in married couple conversation show that in general, lexical cues, manually transcribed or derived from noisy ASR lattice, were most effective for this specific behavior (*e.g.*, around 90%/78% accuracies respectively for estimating *Blame* behaviors in the binary classification task) [46]. The estimation accuracies for the same code just using acoustic [34] and visual cues [16] were close (around 70% accuracies for the same task above). This is not surprising given that language is the most explicit, structured, and controlled method of human communication in expressing blame in a conversation. In general, for across a variety of behavioral codes, multimodal fusion is shown to improve over a single modality [47], [48], [49].

F. Behavioral entrainment

There have been computational studies on underlying behavioral mechanisms. For instance, human behaviors in interactions, such as movement, facial expression, physiology, and emotion, often become alike or coordinated. This phenomenon is called entrainment [11], or in other close terms — mimicry, mirroring, synchrony, *etc.* [50]. Behavior entrainment is an underlying mechanism characterizing rapport and affect between interlocutors [12], [13] and is connected with clinical outcomes in psycho-therapy [14]. Entrainment in behavioral sciences has been modeled largely qualitatively in the past, however recently some initial attempts on computational models of entrainment are emerging [51], [52], [53], [54].

Delaherche *et al.* [55] have conducted a survey of recent works on computational model of human interaction synchrony. In sum, the study of synchrony have spanned multimodal signal forms, and have been evaluated against human annotation, or through comparison between true interaction and *pseudo*-interaction. The synchrony measures of two interlocutors could be mainly categorized into three classes: correlation of multimodal features; derived from phase or spectrum of the signals; and derived from recurring instances from each person's "bag of features".

III. Modeling kinemes of head motion

We have introduced kineme in Sec. I as the elementary unit of motion in the kinesics theory. In this section, we model the kinemes of head motion in a data driven way, beginning with head motion estimation and segmentation.

A. Motion estimation

In this work, we compute optical flow of the face region as an estimate of motion. This is also used in other work such as by Martin *et al.* [56]. The procedure described here is simple and effective for the data we are analyzing: the audiovisual data is of relatively low-quality and the participants are seated and thus their lower body is relatively immobile. More advanced techniques can be applied, *e.g.*, extended Kalman filter and particle filtering [23], [21]; however, that is not the main focus of this work.

We use the Haar-based cascade classifier implemented in OpenCV [57] to detect the subject's face in each frame, and approximate the face size with the side length S of the square that is marking the detected face. We find the histogram of detected face sizes in integer value bins, and smooth the histogram by averaging with two bins on the left and right, respectively (*i.e.*, 5-point moving average). We estimate the true face size as the mode of the smoothed histogram, *i.e.*, the most likely value, denoted $\hat{S}$. To reduce noise from erroneous detections, we exclude outliers in size by rejecting faces with size $S > 1.2\hat{S}$ or $S < 0.8\hat{S}$. For outliers in location, we estimate the average center of the face position, $\hat{H}_c$, by averaging all detected face center coordinates. After finding the distance from each detected face center H_c to the estimated center of face position $\hat{H}_c$, *i.e.*, $|H_c - \hat{H}_c|$, we exclude face detections with such distance larger than the head size on the horizontal axis, *i.e.*, $|H_{cx} - \hat{H}_{cx}| > \hat{S}$, and larger than half the size of the head on the vertical axis, *i.e.*, $|H_{cy} - \hat{H}_{cy}| > 0.5\hat{S}$.

The above approach is based on the assumption of steady seated posture of the subject, and the thresholds are empirically chosen through observing the video and the subject movements. The frames that have no acceptable face detections are assigned values through linear interpolation of the closest adjacent annotated frames. In Fig. 2 we illustrate some examples of outlier removal in face detection. In each column, we display the distribution of the face positions for one session before and after the outlier removal and interpolation. We can see that the outliers (*e.g.*, near image boundaries) are effectively removed, while the procedure does not influence sessions without outliers. Note that removing outlier face detections does not imply rejection of large head motion, since head motions are computed through optical flows as described below, rather than through changes of face positions.

We compute pixel-wise optical flow in the 2D plane using Farnebäck's algorithm [58], denoted $\vec{\theta}(x, y)$ for pixel at coordinate (x, y). In order to get frame level motion velocities, we propose four types of operators on the field of optical flows over the detected face region, *i.e.*, horizontal, vertical, radial and rotational, denoted $\vec{O_X}(x, y)$, $\vec{O_Y}(x, y)$, $\vec{O_Z}(x, y)$ and $\vec{O_R}(x, y)$, as shown in Fig. 3. The velocity M_W is derived as in (1).

$$M_W = \sum_{x,y} \vec{\theta}(x, y) \cdot \vec{O}_W(x, y), W \in \{X, Y, Z, R\} \quad (1)$$

B. Kinesis activity detection

The aim of Kinesis Activity Detection (KAD) is to separate motion from non-motion in time domain, so that we can focus on motion segments afterwards. We use the motion velocities M_X and M_Y as features to perform the segmentation. We observed that M_Z and M_R were of low SNR and provided little discriminative power thus we did not use them in this study.

We assume motion and non-motion states can each be modeled by a multi-variate Gaussian distribution. We use a 2-component Gaussian Mixture Model (GMM) to describe the distribution of M_X and M_Y. To model transitions from motion to non-motion states we use a 2-state HMM. The GMM and HMM are optimized iteratively. Initially, let $M(t) = \sqrt{M_X^2(t) + M_Y^2(t)}$ be the total velocity, where $t = 1, 2, \cdots, T$ is the time index and T is the total duration. We assign the top 20% samples of $M(t)$ to the motion class, and the rest to the non-motion class, for an Expectation-Maximization (EM) training of GMM. We initialize the self-transition probability of both HMM states to 0.9. The state of each sample is estimated in Maximum-a-Posteriori (MAP) sense using Forward-Backward algorithm on the HMM. Then the GMM is re-trained based on the new state estimation, while the HMM parameters are also updated. We repeat this process for 30 iterations, which is sufficient for convergence for our experimental data. In the end, we obtain the optimized state estimation.

Moreover, if a pause less than 0.2 seconds exists between two motion segments each longer than 1 second, we consider the pause as part of the motion and merge it with neighboring motion segments. We then remove short motion segments that are less than 1 second. The smoothing is applied to ensure smoothness and noise reduction and that segments are of significant size to include meaningful behavioral gestures.

C. Motion normalization and segmentation

Real world data recordings, such as the ones we are dealing with, tend to have large variability in their acquisition settings, methods, equipment, lightning, subject characteristics, *etc.* Hence the acquired data require normalization to overcome the variation, and to generalize across subjects.

In this study we apply two steps of normalization:

First, we observed that due to differences in the sitting posture of the subjects such as leaning on the couch or sitting upright in a chair, the main directions of head motion might not align with the horizontal and vertical image axes. To correct for this alignment issue we rotate the extracted motion M_X and M_Y along the main motion directions. These are the Principal Component Analysis (PCA) components in the 2D plane of M_X and M_Y, based on the motion segments estimated in the KAD step. We then project the motion vector (M_X, M_Y) onto the new PCA axes, resulting in $(M_X^{'}, M_Y^{'})$.

Second, we observed that the amplitude of extracted head motion varied among subjects, due to the distance of camera to the head, and the nature of inter-person heterogeneity. To correct for this we apply a zero-mean unit-variance normalization separately on $M_X^{'}$ and $M_Y^{'}$ for each session, obtaining $\tilde{M}_X$ and $\tilde{M}_Y$.

KAD provides the estimation of motion states, however, there may be very long sequences of motion. Human head motion may change rapidly, and we assume that those long sequences contain multiple kinemes which are short-time stationary. We thus apply a shifting window on long motion segments to localize the motion events. Motion segments longer than 3 seconds are broken into 2-second motion events with 1 second overlap; otherwise preserved as a single motion event. As a result, the windowed motion events vary in length between 1 second to 3 seconds, due to KAD and windowing, with the bulk of the events being of 2 seconds.

D. Representation of head motion

We use Linear Predictive (LP) models to extract parametric representation of motion events, assuming head motion can be approximated by an auto-regressive process. Specifically, LP models are expected to be effective in capturing the frequency of repeated movements such as nodding and shaking. LP derived features are also consistent in dimension for motion events of varying lengths. In the end we convert the LP coefficients to Line Spectral Frequencies (LSF) since those exhibit better quantization properties [59].

We compute the LSF for $\tilde{M}_X$ and $\tilde{M}_Y$, then concatenate the results. Let $L_j^i(x)=(l_x(1), l_x(2), \cdots, l_x(N))$ be the order-*N* LSF feature, extracted from the segment of $\tilde{M}_X$ corresponding to the *i*-th motion events in the *j*-th stream. Let $L_j^i(y)=(l_y(1), l_y(2), \cdots, l_y(N))$ be the order-*N* LSF feature extracted from the corresponding segment of $\tilde{M}_Y$. We denote $L_j^i=(L_j^i(x), L_j^i(y))$ as the 2*N* -dimensional feature vector representing the motion events.

E. Generalization of head motion types

To generalize our model for various types of head motion, we employ a Gaussian Mixture Model (GMM) over the LSF features [60]. We assume that the mixture model is an approximation to the set of kinemes, where each component may be associated with one kineme, *i.e.*, one type of head motion. The GMM can also be viewed as a soft clustering approach compared to a hard-labeled clustering such as K-means. The posterior of a motion event evaluated on each component can be interpreted as a partial cluster membership. Such a relaxation allows modeling of ambiguous motion events as combinations of multiple motion prototypes.

The GMM training is conducted on LSF features extracted from all motion events in all training sessions. We use the K-means algorithm with K clusters to initialize the training, and use the Expectation-Maximization algorithm for learning the parameters of the GMM. As a result, we obtain the prior probabilities π_k, the mean vectors $\mu_k = (\mu_k(1), \mu_k(2), \cdots, \mu_k(2N))$, and the variance vectors $\sigma_k = (\sigma_k(1), \sigma_k(2), \cdots, \sigma_k(2N))$, *i.e.*, the diagonal of the assumed diagonal covariance matrix corresponding to the feature vector of dimension $2N$, where $k = 1, 2, \cdots, K$ is the component index.

For GMM inference, let us consider L_j^i to be the feature vector. The likelihoods $P(L_j^i|k)$ and posteriors $P(k|L_j^i)$, $k = 1, 2, \cdots, K$, are derived from the GMM [60].

We accumulate the posteriors of all the motion events for a subject in one session, and consider it as a soft histogram of motion types. Such aggregated partial counts may reflect the composition of motion activities of a subject. Let the vector of posterior sum be F_j, the session duration be T_j, we obtain F_j in (2)

$$F_j(k) = \frac{1}{T_j} \sum_i P(k|L_j^i), k = 1 \cdots K \quad (2)$$

Note that the objective of EM is to maximize the likelihood function. Since this is an unsupervised optimization this may not be the same as optimizing with respect to behaviorally meaningful motion types. Due to the latent nature of motion types and its relation to behavioral characteristics, one way of optimizing the GMM may be selecting the model based on its discriminative power towards behavior code prediction (our intended application). In practice, we train an ensemble of GMMs that are initialized with different K-means derived parameters. With limited data, we do not have a distinct, large, held out set for the selection of optimal GMM that would eliminate the need for a GMM ensemble. Therefore we make use of the ensemble in two ways: (1) by integrating the results from each individual model, (2) by employing one sample of the training set as a development set to choose the best performing model (this will lead to $X \cdot (X - 1)$ -fold validation for X distinct samples).

To summarize, in this section we have modeled kinemes of head motion in three conceptual steps: segmentation, representation, and generalization of motion events. These steps are shown in Fig. 4.

IV. Head motion similarity in interaction

A. Modeling head motion similarity

Our goal for head motion similarity modeling is to find similar motion activities from the two interlocutors, based on the proposed head motion model, and define the metric of similarity both for comparing motion events and for comparing the overall behavior during a time interval.

Let L_w^i and $L_h^{i'}$ be the feature vectors of two head motion events from an interacting dyad, say the subject *w* (wife) and *h* (husband), respectively. Let the head motion be modeled using the GMM in Sec. III-E, and the posterior of feature vector of head motion be denoted $P(k|L)$ for component *k*. Mathematically, we employ a divergence formulation to quantify similarity, where lower divergence is equivalent to higher similarity, and *vice versa*. Let the symmetric Kullback-Leibler (KL) divergence of two posterior distributions be denoted as in (3).

$$\mathrm{KL}(\mathrm{L}_\mathrm{w}^\mathrm{i}, \mathrm{L}_\mathrm{h}^{\mathrm{i}'}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{P}(\mathrm{k}|\mathrm{L}_\mathrm{w}^\mathrm{i}) \log \frac{\mathrm{P}(\mathrm{k}|\mathrm{L}_\mathrm{w}^\mathrm{i})}{\mathrm{P}(\mathrm{k}|\mathrm{L}_\mathrm{h}^{\mathrm{i}'})} + \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{P}(\mathrm{k}|\mathrm{L}_\mathrm{h}^{\mathrm{i}'}) \log \frac{\mathrm{P}(\mathrm{k}|\mathrm{L}_\mathrm{h}^{\mathrm{i}'})}{\mathrm{P}(\mathrm{k}|\mathrm{L}_\mathrm{w}^\mathrm{i})} \quad (3)$$

To avoid numerical instability caused by zero values in the posterior, we add a small positive value $\varepsilon = 1 \times 10^{-5}$ to all elements of the posterior distribution and re-normalize before substituting to (3).

Let $\mathscr{B}_w = \{L_w^i\}_{i=1}^I$ be the set of feature vectors corresponding to the head motion events of subject *w* during a certain time interval $\mathscr{T}$, where *I* is the total count of motion events.

Similarly for subject *h*, denote $\mathscr{B}_h = \{L_h^{i'}\}_{i'=1}^{I'}$ with respect to the same time interval $\mathscr{T}$. We define the similarity measure for subject *w* and *h* over $\mathscr{T}$ in the form of divergence as follows.

1. Compute pairwise symmetric KL divergence for all pairs of motion events in $\mathscr{B}_w$ and $\mathscr{B}_h$, resulting in a matrix *DIV* of size $I \times I'$, where $DIV(i, i') = \mathrm{KL}(\mathrm{L}_\mathrm{w}^\mathrm{i}, \mathrm{L}_\mathrm{h}^{\mathrm{i}'})$.
2. Convert the matrix *DIV* to a vector $D = (d(1), d(2), \cdots, d(II'))$, sorted in ascending order.
3. Compute similarity measures $\mathrm{div}(\mathscr{B}_w, \mathscr{B}_h, \rho)$ per (4), where $0 < \rho < 1$ is a tuning parameter, and $\lfloor \rho II' \rfloor$ is the count of elements in *D* that is averaged.

$$\mathrm{div}(\mathscr{B}_\mathrm{w}, \mathscr{B}_\mathrm{h}, \rho) = \frac{1}{\lfloor \rho \mathrm{II}' \rfloor} \sum_{\mathrm{n}=1}^{\lfloor \rho \mathrm{II}' \rfloor} \mathrm{d}(\mathrm{n}) \quad (4)$$

We capture very similar pairs of motion events from two interactants with $\mathrm{div}(\mathscr{B}_w, \mathscr{B}_h, \rho)$. This focus on extreme values in *D* is based on the intuition that participants' very similar motions are more behaviorally meaningful (salient). For simplicity in computation, we do

not match motion events of two subjects one-to-one, but take the mean of pairwise symmetric KL divergence as an averaged measure.

B. Dynamics of motion similarity

The motion similarity measure is influenced by the animation degree of the interlocutors. Without any normalization, the more animated the couple, the more likely they are to exhibit similar motion events. Limited by the duration of interaction sessions in our data, we divide each session into two halves and employ the first half as a normalizing factor thus enabling us to evaluate the, now, normalized degree of similarity change along the session.

Let $\mathscr{B}_w^1$ and $\mathscr{B}_w^2$ be the sets of motion events for subject *w* in the first and second halves of the interaction, respectively. Similarly let $\mathscr{B}_h^1$ and $\mathscr{B}_h^2$ be that for subject *h*. Let the relative change of similarity measure derived from the GMM with index *m* be denoted $\tilde{R}_m(w,h)$, as in (5).

$$\tilde{R}_m(w,h)=\frac{\mathtt{div}(\mathscr{B}_{\mathtt{w}}^2,\mathscr{B}_{\mathtt{h}}^2,\rho)}{\mathtt{div}(\mathscr{B}_{\mathtt{w}}^1,\mathscr{B}_{\mathtt{h}}^1,\rho)} \quad (5)$$

Our goal is to establish the usefulness of this measure to characterize behavioral synchrony or entrainment. Since we do not have access to direct ground truth measures of entrainment for validation, we do this indirectly by examining how well the proposed similarity measures will statistically explain behavioral constructs (codes) where entrainment is implicated. We use the averaged similarity based on the ensemble of GMMs as a more robust measure, due to a lack of held out dataset for model selection as we discussed in Section. III-E. Let the log-scale averaged relative change of similarity measure be denoted *R*(*w*, *h*), as in (6). We use log operation to stretch the scales below and above 1 before correlating to behavior codes. *R*(*h*, *w*) is equal to *R*(*w*, *h*) since we employ symmetric KL divergence.

$$R(w,h)=\log\left(\frac{1}{M}\sum_{m=1}^{M}\tilde{R}_m(w,h)\right) \quad (6)$$

V. Couple therapy corpus

A. Data collection and annotation

The corpus used in this study comprises audio-visual recordings of seriously and chronically distressed couples in dyadic conversations addressing a problem in their marriage. The data were collected by the University of California, Los Angeles and the University of Washington [61]. Each couple talked about two separate problems, one chosen by the wife and one by the husband, for 10 minutes each. These discussions took place at three points in time during the therapy process: before the psychotherapy began, 26 weeks into the therapy and 2 years after the therapy session finished. The full database is 96 hours long and contained 574 sessions. The video format was 704 × 480 pixels, 30 fps, with a screen split and one spouse on each side.

Both spouses in all sessions were evaluated individually following two expert designed coding systems, the Couples Interaction Rating System 2 (CIRS2) [62] and the Social Support Interaction Rating System (SSIRS) [63]. The CIRS2 contained 13 behavior codes and was specifically designed for conversations involving a problem in relationship, while the SSIRS consisted of 20 codes that measured the emotional component of the interaction and the topic of conversation. These codes serve the interests of various research and clinical questions. Of these, for this work, we focus on four codes — *Acceptance*, *Blame* from CIRS2, and *Positive*, *Negative* from SSIRS[1], which have relations to behavioral entrainment that is grounded in theoretical studies [64], [65], [66]. Moreover, these also reflect general affect and attitudes of the subjects. The remaining codes available in the full database, which are associated with diverse research questions, are not conceptually relevant or directly within the scope of the current study's focus on head-motion based behavioral modeling.

The researchers trained a group of undergraduate students majoring in psychology to perform the annotation; the annotators acquired adequate knowledge in the domain and could be considered as "experts" compared to naive coders. At least three students were assigned to the same session, where they would watch the entire session and give an overall score on each code. Each score is on a discrete numerical range from 1 to 9. We use the average score among coders as ground truth. Note that the codes only measure how much a particular behavior of interest occurs, independent of how much their opposite occurs. For example, both *Positive* and *Negative* codes can have high value if they are both present in the interaction.

B. Data pre-processing and filtering

The video quality of the recordings (that took place across several different clinical settings) is not ideal. There was no calibration carried out, and relative positions of subjects as well as of the cameras were not available as the database was intended originally for human-driven analysis. Therefore, we apply a pre-filtering step to all sessions on the left and right split screen content of the video in order to filter out unqualified sessions. First, we run an OpenCV [57] face detector on one frame per second of the video. Second, the face size is estimated by the mode of the distribution of detected size of the face blocks. Third, we retain sessions which have a face detected on more than 70% of the sampled frames, and have the estimated face size between 80 pixels and 160 pixels, *i.e.*, $\frac{1}{6}$ to $\frac{1}{3}$ of the image height. In these recordings, the upper body of a subject is generally present while it is uncertain if the hands are captured.

We constructed two subsets of the dataset for the two analyses of this work. First, we considered subjects on an individual basis, and collected samples from only one side or both sides of the split screen. This resulted in 561 data sequences of a single subject each. We denote this collection of data as $\mathscr{D}_1$.

[1]*Acceptance* indicates understanding and acceptance of partner's views, feelings and behaviors. *Blame* indicates that one blames, accuses, or criticizes the partner, uses critical sarcasm, makes character assassinations. *Positive* and *Negative* are overall rating of the positive and negative affect the target spouse showed during the interaction. Examples include overt expressions of warmth, support, acceptance, affection, positive negotiation and compromise for *Positive*, and rejection, defensiveness, blaming, and anger for *Negative*.

Second, we considered the interaction of two interlocutors, *i.e.*, a couple, and collected sessions in which the video quality of both partners satisfy our pre-filtering conditions. As a result 163 sessions (326 single-subject sequences, 66 unique pairs of couples) were obtained. We denote this collection of data as $\mathscr{D}_2$. Compared to our previous works [16], [17], we now utilize more than twice the data.

VI. Experiments and hypothesis tests

A. Inferring expert judgment on behavior using head motion signal

We address a problem of classifying "low" *vs.* "high" presence of certain behavior codes for a subject in this section, in order to investigate the effectiveness of the proposed head motion model. While the behavior codes are assigned discrete values from 1 to 9 by the human coders, we divide the dataset $\mathscr{D}_1$ for each code into three parts: $\mathscr{D}_1^+$ and $\mathscr{D}_1^-$ for the top 25% of sessions with the highest scores and the 25% of sessions with the lowest scores, respectively; as well as $\mathscr{D}_1^0$ for the middle 50% of the sessions.

We consider four behavior codes: *Acceptance, Blame, Positive* and *Negative* which are introduced in Sec. V-A. The human experts achieved 0.7 correlation in their annotations of these codes, pointing to high inter-coder agreement. This suggests that the behaviors are consistently represented and perceived by human coders and thus provide a meaningful challenge to address using computational methods.

In contrast to [16] where the model was trained on $\mathscr{D}_1^+$ and $\mathscr{D}_1^-$, in this work we train our model on $\mathscr{D}_1^0$. This keeps training and testing data disjoint thus it avoids a need for cross-validation on the model construction.[2] For the linear predictive analysis step (Sec. III-D), a higher filter order N yields a better fit and decreased residual error but at the cost of overfitting and increasing the feature dimensions and hence the complexity of the resulting GMM. For our experiments, we choose a linear predictive analysis order of $N = 10$ based on pilot trials, where any higher order does not improve the performance. The number of GMM mixtures K should be selected to reflect the number of head motion kinemes. However, the number of kinemes is unknown given that the nature of head motion structure is not obvious, so we experimentally examine K from 3 to 25. The lower bound is chosen to model a very coarse category of kinemes, while the upper bound is much larger than the number of head motion types proposed in most existing coding systems. This larger upper limit is also motivated by the assumption that automatic clustering may capture finer types than manual labeling, although in practice it is constrained by available data samples to robustly learn the GMM. We run the GMM training with randomized initialization for $M = 50$ times to get an ensemble of GMMs, where the number 50 is chosen as a trade-off between having an adequate number to increase robustness and assuring affordability in terms of computational complexity.

We test the model by setting up a binary classification problem where the samples in $\mathscr{D}_1^+$ and $\mathscr{D}_1^-$ are in two classes. The features for training the binary classifiers are the

[2]Cross validation is still needed for other aspects of the analysis as described below.

accumulated posterior vectors of the motion events, defined as F_j in Sec. III-E. We use a linear Support Vector Machine (SVM) as our binary classifier [67]. Limited by the number of samples, *i.e.*, 282 in total combining $\mathscr{D}_1^+$ and $\mathscr{D}_1^-$, we carry out a leave-one-subject-out cross validation. This means we train the SVM on all but one subject, and test on that subject; and we repeat the process for every subject in the dataset for classification. Constrained by data sparsity, we use this scheme to validate our method's ability to generalize across different subjects. The total number of subjects (wives and husbands) in experiments with respect to each code is included in Table I.

We consider two approaches for fusing the M models given the same K. First, we choose the classifier that performs best on the training data itself in each round, denoting its test performance as Acc_t. This in a sense is using the training data as the development data, due to data sparsity. The second approach is to predict using all the M GMMs and employ majority-voting as the decision, with the test performance denoted Acc_e. The assumption here is that the clustering that converges at other local minima is detecting events independent to the behavior codes of interest, hence averaging would cancel those out.

We plot the results of Acc_t and Acc_e for the four target behavior codes (*Acceptance*, *Blame*, *Positive*, *Negative*) and different values of K in Fig. 5. As we can see, the best results of both Acc_t and Acc_e exceed 60% accuracy, and are close to 70% for some codes. However, the best-performing number of clusters K differs by codes. There is a general trend of better accuracy with larger K, but the fluctuation of accuracy against K is also high. This means that increasing the number of clusters by one may lead to better or worse results. Since the space of motion types is latent, it is difficult to answer what is the optimal K from a theoretical perspective.

Following the above discussion, we want to examine the feasibility of automatic model selection among all K values and ensembles of GMMs, which was not addressed previously in [16]. Due to the lack of a development set, we run an innerlayer of cross-validation against all GMMs, *i.e.*, during each round of testing, among the X-but-one training subjects, we again repeatedly train on X-but-two subjects and test on the corresponding left out subject for $X - 1$ times. This leads to a total of $X \cdot (X - 1)$ times of training and testing binary classifiers. We evaluate the effectiveness of GMMs by the average accuracy in $X - 1$ times of cross-validation. Based on the evaluated performance, we may select a group of models in practice. In the experiment, we check the averaged test accuracy Acc_υ based on the majority vote of the selected top V models out of 1150 GMMs in each round, where V is an odd number from 1 to 199. In Table I we show Acc_υ for $V = 199$, as we find that Acc_υ converges around $V = 99$ and in general stays stable afterwards. The values of Acc_υ are significantly better than chance (accuracy of 0.5) at $p \leq 0.01$ based on one-tailed binomial test.

B. Hypothesis tests on motion similarity dynamics

We examine the hypothesis that the relative change of similarity measure is linked to target behavioral characteristics of subjects, based on dataset $\mathscr{D}_2$ containing 163 sessions. Denote $\mathscr{R}$ as the vector of similarity measures for both the wives and the husbands in (7):

$$\mathscr{R}=(R(\mathtt{w}_1,\mathtt{h}_1),R(\mathtt{h}_1,\mathtt{w}_1),R(\mathtt{w}_2,\mathtt{h}_2),R(\mathtt{h}_2,\mathtt{w}_2),\cdots,\ R(\mathtt{w}_{163},\mathtt{h}_{163}),R(\mathtt{h}_{163},\mathtt{w}_{163})) \quad (7)$$

Let $Y(w)$, $Y(h)$ be the behavior code value for subject w and h, respectively. Denote $\mathscr{Y}$ as the vector of a certain behavior code for all the subjects as in (8).

$$\mathscr{Y}=(Y(\mathtt{w}_1),Y(\mathtt{h}_1),Y(\mathtt{w}_2),Y(\mathtt{h}_2),\cdots,Y(\mathtt{w}_{163}),Y(\mathtt{h}_{163})) \quad (8)$$

We investigate the Pearson's correlation coefficients between $\mathscr{R}$ and $\mathscr{Y}$, with hypotheses testing performed using student's t-distribution. The null and alternative hypotheses are:

$\mathbf{H}_0$ $\mathscr{R}$ and $\mathscr{Y}$ are uncorrelated.

$\mathbf{H}_a$ there is some correlation between $\mathscr{R}$ and $\mathscr{Y}$.

In [17] we trained the head motion model on the same set for testing correlation. In this work, we use the set $\mathscr{D}_c = \{s|s \in \mathscr{D}_1, s \notin \mathscr{D}_2\}$ to train the head motion model, which is disjoint to $\mathscr{D}_2$. Similar to the previous experiment, we set $N = 10$, $M = 50$, and $K \in \{3, 4, \cdots, 25\}$. We sample the value of parameter ρ ($0 < \rho < 1$) including 0.01, 0.02, and from 0.05 to 0.25 with a step size of 0.05. We found that there were sessions with extreme values of $r \in \mathscr{R}$, where the estimation of head motion was contaminated by recording artifacts as the video were originally recorded on tapes. In order to avoid the influence of outliers, we exclude sessions of the top 3% largest $|r|$ values (on two tails in logarithm domain) for each test of correlation.

Moreover, to verify that any found correlations are meaningful effects of the interaction, we conduct the correlation analysis with random pairings of interacting subjects. In other words, the similarity and relative change are computed based on subjects who did not interact with each other (*i.e.*, not "true" couples). We repeat the shuffling for 100 times.

As a result, we find consistent significant correlations for the four behavior codes of interest: *Acceptance*, *Blame*, *Positive*, and *Negative*. In Fig. 6a we show the correlations obtained with $\rho = 0.05$. In general, we can see that the signs of the correlations confirm to the polarity of codes. For example, negative emotional codes such as *Blame* and *Negative* are positively correlated with divergence, and *vice versa* for positive emotional codes *Acceptance* and *Positive*.

The correlations are significant at $p < 0.05$ for K in the range of 7 to 9, and also 13 to 16 for *Positive* and *Negative*. For the above cases, we reject $\mathbf{H}_0$ and suggest $\mathbf{H}_a$. The most prominent correlations are obtained with $K = 9$. Compared with the findings in Sec. VI-A, we see that larger values of K may provide higher discriminative power; however, a very fine clustering of motion types may not help tracking the similarity of motion.

In addition, we display the correlation with $K = 9$ and different ρ values in Fig. 6b. We see that smaller ρ yields stronger correlation. The results are comparable for $\rho = 0.02$ and $\rho = 0.05$, which suggests that the setting of ρ is not necessarily fixed to a particular value but robust over a range. The correlation decreases as ρ becomes larger. This might suggest that closer matching of behaviors are more salient, *i.e.*, pairs of motion with small divergence are more influential on the coder's judgment of the interlocutor's behavioral characteristics.

In Table II we show the statistics of the correlation results in randomized pairings of subjects, with $K = 9$ and $\rho = 0.05$. The mean of correlation is close to 0, and any significant correlation ($p < 0.05$) is beyond at least one standard deviation. This suggests that although spurious high correlation may occur in the random pairings, it does not generally happen.

In sum, we have shown the relation between the proposed head motion similarity measure and the expert-specified behavioral characteristics. In the future we would like to investigate if the relation can be strengthened such that the similarity can become a feature for code value inference.

VII. Discussion

A. Analysis of the head motion model

Our head motion model is derived by a data driven approach; therefore, we would like to analyze what kind of kinemes the model captures. We take one GMM from the ensemble that is trained with $N = 10$, $K = 9$ on $\mathscr{D}_c$ as an example, which yields strong correlation in Sec. VI-B. Recall that $\mu_1, \mu_2, \cdots, \mu_9$ are the mean vectors in the GMM, each composed by the averaged $L(x)$ and $L(y)$ that are LSF parameters. We convert each $L(x)$ and $L(y)$ to linear predictive coefficients $L'(x)$ and $L'(y)$, which can be viewed as coefficients of auto-regressive filters.

We plot the filter impulse responses along the horizontal and vertical directions for the 9 clusters in Fig. 7. In addition, we retrieve motion events of high posterior probability (> 0.95) on each cluster from $\mathscr{D}_2$, and watch the collection of corresponding video segments per cluster. Associating the impulse responses and the collected motion events, we find that the clustering identifies not only head nod and shake, but also the magnitude (small or large) and speed (slow or fast). Such clustering confirms Hadar's findings [2] that were introduced in Sec. I of the paper. Recall that Hadar proposed five classes of head motion, namely "slow, ordinary, rapid, posture shift, and tremor". We can see there is substantial overlap between the knowledge driven clustering proposed by Hadar and the automatic clustering learned from data and annotations of high level behavioral characteristics. The details of cluster descriptions are included in Fig. 7.

The above result emphasizes that head motion should be studied with the aspects of magnitude and speed taken into consideration. It also shows that linear predictive analysis method is suitable for capturing these properties of head motion. In both experiments, the datasets for training and applying the head motion model are disjoint, showing that the modeling approach generalizes well.

B. Future directions

In the future, there are many directions to advance behavior analysis. From the head motion modeling perspective, direct and more accurate motion acquisition methods such as Kinect along with finer face tracking can be applied to get 3D motion signal. Joint segmentation and clustering algorithms in addition to novel representation methods can be designed to formulate a better set of typical motion exemplars. From the behavior analysis perspective, machine learning methods can be applied to answer how these motion patterns link to the

expert's behavior judgment. Alongside, methods for effectively incorporating domain knowledge into the algorithm should be developed, *e.g.*, characterizing the behavioral meaning of gestures.

Behavior is expressed multimodally; thus modeling of a single modality, *e.g.*, head motion, may provide useful information about particular behavioral characteristics, but may not contain the full information in the multimodal expression. The behavior code classification method in this work based on head motion, though not adequate to work alone, provides an important stream of behavioral cues that may be integrated within a multimodal system for more effective analysis. Future modeling techniques may consider multiple channels jointly.

Moreover, certain parts of the interaction may be more important than others in terms of carrying behavioral meaning, which is often referred as "salient" events [68]. Identifying salient events and weighting the signal derived features may lead to better understanding of the big picture of human interaction. Human interaction is also dynamic and interactive. We simplified the problem by using "bag-of-motion-events" in the experiments, while in the future we would like to model the mutual influence among the interlocutors.

The above research goals have to be supported by sufficient *real* rather than acted interaction data, ideally from natural encounters. Proper data collection systems must be implemented, ideally being multimodal, multi-channel, high quality and non-intrusive. Data annotation is also a key part of the study. So far we have been relying on psychologists to score the behavior along the dimensions of well-designed coding systems. Coder reliability and the consistency among coders are some of the key issues that influence both the modeling and validation. Quantifying the human coder biases may help to better understand the ground-truth. In cases where constraints on data publication due to subject privacy can be relaxed, one might also take advantage of the crowd-sourcing approaches to obtain other non-expert views on the data.

VIII. Conclusion

Head motion is an important channel of nonverbal communication that conveys rich information. In this paper we described a data driven method to model head motion in human interaction, in order to analyze socio-communicative and affective behavioral characteristics of interacting partners. We first segmented the stream of 2D motion into small chunks of motion events, then extracted linear predictive features to represent the signal, and finally constructed GMMs to generalize the motion into kinemes. We applied the model in a binary classification problem of "high" *vs.* "low" presence of certain target behavior codes that were annotated by domain experts. We achieved accuracies around 60% to 70% in general using the proposed model. In addition, we derived a similarity measure of head motion in order to approximate behavioral entrainment in interaction. Through statistical hypotheses tests, we found that the relative change of similarity correlated with behavior code values, where entrainment processes are conceptually implicated to be at work. These results demonstrated the promise of the proposed model.

In the future, we would like to work on improvement of head motion signal extraction, multimodal analysis of behavior, saliency analysis of behavioral cues, as well as extended data collection. We believe signal processing approaches on multimedia observational data can greatly impact the understanding of human behavior and its translation to societal applications such as in health and educational assessment and intervention.

Acknowledgments

This work is supported by NIH, NSF and DoD.

References

1. Harrigan, J.; Rosenthal, R.; Scherer, K. The new handbook of Methods in Nonverbal Behavior Research. New York: Oxford University Press; 2005. p. 137-198.
2. Hadar U, Steiner T, Grant E, Rose FC. Kinematics of head movements accompanying speech during conversation. Human Movement Science. 1983; 2(1):35–46.
3. Hadar U, Steiner T, Grant E, Rose FC. Head movement correlates of juncture and stress at sentence level. Language and Speech. 1983; 26(2):117–129. [PubMed: 6664179]
4. Hadar U, Steiner TJ, Grant EC, Rose FC. The timing of shifts of head postures during conservation. Human Movement Science. 1984; 3(3):237–245.
5. Hadar U, Steiner T, Rose FC. The relationship between head movements and speech dysfluencies. Language and Speech. 1984; 27(4):333–342. [PubMed: 6536844]
6. Hadar U, Steiner T, Clifford Rose F. Head movement during listening turns in conversation. Journal of Nonverbal Behavior. 1985; 9(4):214–228.
7. McClave E. Linguistic functions of head movements in the context of speech. Journal of Pragmatics. 2000; 32(7):855–878.
8. Ekman P, Friesen W. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Nonverbal communication, interaction, and gesture. 1981:57–106.
9. Birdwhistell, R. Kinesics and context: essays on body motion communication. Vol. 2. University of Pennsylvania Press; 1970.
10. Bousmalis K, Mehu M, Pantic M. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. Image and Vision Computing. 2013; 31(2):203–221.
11. Wheatley T, Kang O, Parkinson C, Looser C. From mind perception to mental connection: Synchrony as a mechanism for social understanding. Social and Personality Psychology Compass. 2012; 6(8):589–606.
12. Bernieri F. Coordinated movement and rapport in teacher-student interactions. Journal of Nonverbal behavior. 1988; 12(2):120–138.
13. LaFrance M. Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. Social Psychology Quarterly. 1979:66–70.
14. Ramseyer F, Tschacher W. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. Journal of consulting and clinical psychology. 2011; 79(3):284. [PubMed: 21639608]
15. Verhofstadt LL, Buysse A, Ickes W, Davis M, Devoldre I. Support provision in marriage: the role of emotional similarity and empathic accuracy. Emotion. 2008; 8(6):792. [PubMed: 19102590]
16. Xiao B, Georgiou P, Baucom B, Narayanan S. Data driven modeling of head motion towards analysis of behaviors in couple interactions. Proc. ICASSP. 2013
17. Xiao B, Georgiou P, Lee C, Baucom B, Narayanan S. Head motion synchrony and its correlation to affectivity in dyadic interactions. Proc. ICME. 2013
18. Murphy-Chutorian E, Trivedi MM. Head pose estimation in computer vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009; 31(4):607–626. [PubMed: 19229078]

19. Zhao, G.; Chen, L.; Song, J.; Chen, G. Large head movement tracking using sift-based registration. Proceedings of the 15th international conference on Multimedia; ACM; 2007. p. 807-810.
20. Kiruluta A, Eizenman E, Pasupathy S. Predictive head movement tracking using a kalman filter. IEEE Transactions on Systems, Man, and Cybernetics, Part B. Cybernetics. 1997; 27(2):326–331.
21. Miao, Y-Q.; Fieguth, P.; Kamel, MS. Image Analysis and Recognition. Springer; 2011. Maneuvering head motion tracking by coarse-to-fine particle filter; p. 385-394.
22. Li P, Zhang T, Ma B. Unscented kalman filter for visual curve tracking. Image and Vision Computing. 2004; 22(2):157–164.
23. Ababsa, F.; Mallem, M.; Roussel, D. Proc. ICRA. Vol. 1. IEEE; 2004. Comparison between particle filter approach and kalman filter-based technique for head tracking in augmented reality systems; p. 1021-1026.
24. Kondori, FA.; Yousefi, S.; Li, H.; Sonning, S. Proc. WCSP. IEEE; 2011. 3D head pose estimation using the kinect; p. 1-4.
25. Fanelli, G.; Gall, J.; Van Gool, L. Proc. ISCCSP. IEEE; 2012. Real time 3D head pose estimation: Recent achievements and future challenges; p. 1-4.
26. Kapoor, A.; Picard, RW. Proceedings of the 2001 workshop on Perceptive user interfaces. ACM; 2001. A real-time head nod and shake detector; p. 1-5.
27. Tan W, Rong G. A real-time head nod and shake detector using hmms. Expert Systems with Applications. 2003; 25(3):461–466.
28. Kang, YG.; Joo, HJ.; Rhee, PK. Knowledge-Based Intelligent Information and Engineering Systems. Springer; 2006. Real time head nod and shake detection using hmms; p. 707-714.
29. Nguyen, L.; Odobez, J-M.; Gatica-Perez, D. Proc. ICMI. ACM; 2012. Using self-context for multimodal detection of head nods in face-to-face interactions; p. 289-292.
30. Morency L-P, Sidner C, Lee C, Darrell T. Head gestures for perceptual interfaces: The role of context in improving recognition. Artificial Intelligence. 2007; 171(8):568–585.
31. Akakın HÇ, Sankur B. Robust classification of face and head gestures in video. Image and Vision Computing. 2011; 29(7):470–483.
32. Busso C, Deng Z, Grimm M, Neumann U, Narayanan S. Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Transactions on Audio, Speech, and Language Processing. 2007; 15(3):1075–1086.
33. Narayanan S, Georgiou P. Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proceeding of IEEE. 2013; 101(5):1203–1233.
34. Black MP, Katsamanis A, Baucom BR, Lee C-C, Lammert AC, Christensen A, Georgiou PG, Narayanan SS. Toward automating a human behavioral coding system for married couples interactions using speech acoustic features. Speech Communication. 2013; 55(1):1–21.
35. Can D, Georgiou P, Atkins D, Narayanan S. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. Proc. InterSpeech. 2012
36. Xiao, B.; Can, D.; Georgiou, PG.; Atkins, D.; Narayanan, SS. Asia-Pacific Signal & Information Processing Association Annual Summit and Conference. IEEE; 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy.
37. Bone D, Black MP, Lee C-C, Williams ME, Levitt P, Lee S, Narayanan S. Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. Proc. Interspeech. 2012
38. Chaspari T, Lee C-C, Narayanan SS. Interplay between verbal response latency and physiology of children with autism during eca interactions. Proc. InterSpeech. 2012 Sep.
39. Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schröder M. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing. 2012; 3(1):69–87.
40. Kendon A, Sigman S. Commemorative essay. Ray L. Birdwhistell (1918–1994). Semiotica. 1996; 112(3–4):231–262.
41. Poppe R. A survey on vision-based human action recognition. Image and vision computing. 2010; 28(6):976–990.

42. Niebles JC, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision. 2008; 79(3):299–318.
43. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical dirichlet processes. Journal of the American Statistical Association. 2006; 101(476)
44. Fox EB, Sudderth EB, Jordan MI, Willsky AS. A sticky hdp-hmm with application to speaker diarization. The Annals of Applied Statistics. 2011; 5(2A):1020–1056.
45. Sargin ME, Yemez Y, Erzin E, Tekalp AM. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008; 30(8):1330–1345. [PubMed: 18566489]
46. Georgiou, P.; Black, M.; Lammert, A.; Baucom, B.; Narayanan, S. Proc. ACII. Springer; 2011. "that's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?; p. 87-96.
47. Katsamanis, A.; Gibson, J.; Black, MP.; Narayanan, SS. Proc. ACII. Springer; 2011. Multiple instance learning for classification of human behavior observations; p. 145-154.
48. Gibson J, Xiao B, Georgiou PG, Narayanan SS. An audiovisual approach to learning salient behaviors in couples' problem solving discussions. Proc. ICME. 2013 Jul.
49. Black MP, Georgiou PG, Katsamanis A, Baucom BR, Narayanan S. "you made me do it": Classification of blame in married couples' interactions by fusing automatically derived speech and language information. Proc. Interspeech. 2011
50. Chartrand T, van Baaren R. Human mimicry. Advances in experimental social psychology. 2009; 41:219–274.
51. Varni G, Volpe G, Camurri A. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Transactions on Multimedia. 2010; 12(6):576–590.
52. Sun, X.; Truong, K.; Pantic, M.; Nijholt, A. Proc. SMC. IEEE; 2011. Towards visual and vocal mimicry recognition in human-human interactions; p. 367-373.
53. Lee C-C, Katsamanis A, Black MP, Baucom B, Christensen A, Georgiou PG, Narayanan SS. Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions. Computer, Speech, and Language. 2014; 28(2): 518–539.
54. Xiao B, Georgiou PG, Imel ZE, Atkins D, Narayanan SS. Modeling therapist empathy and vocal entrainment in drug addiction counseling. Proc. of InterSpeech. 2013 Aug.
55. Delaherche E, Chetouani M, Mahdhaoui A, Saint-Georges C, Viaux S, Cohen D. Interpersonal synchrony: A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing. 2012; 3(3):349–365.
56. Martin, S.; Tran, C.; Trivedi, M. Proc. ICPR. IEEE; 2012. Optical flow based head movement and gesture analyzer (ohmega); p. 605-608.
57. Bradski G. The OpenCV Library. Dr. Dobb's Journal of Software Tools. 2000
58. Farnebäck G. Two-frame motion estimation based on polynomial expansion. Image Analysis. 2003:363–370.
59. Kabal P, Ramachandran R. The computation of line spectral frequencies using chebyshev polynomials. IEEE Transactions on Acoustics, Speech and Signal Processing. 1986; 34(6):1419–1426.
60. Bishop, CM. Pattern recognition and machine learning. Springer; 2006. p. 430-439.
61. Christensen A, Atkins D, Berns S, Wheeler J, Baucom D, Simpson L. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. Journal of consulting and clinical psychology. 2004; 72(2):176–191. [PubMed: 15065953]
62. Heavey C, Gill D, Christensen A. Couples interaction rating system 2 (CIRS2). University of California, Los Angeles. 2002
63. Jones J, Christensen A. Couples interaction study: Social support interaction rating system. University of California, Los Angeles. 1998
64. Butner J, Diamond LM, Hicks AM. Attachment style and two forms of affect coregulation between romantic partners. Personal Relationships. 2007; 14(3):431–455.

65. Reed RG, Randall AK, Post JH, Butler EA. Partner influence and in-phase versus anti-phase physiological linkage in romantic couples. International Journal of Psychophysiology. 2013; 88(3): 309–316. [PubMed: 22922526]

66. Ramseyer F, Tschacher W. Synchrony: A core concept for a constructivist approach to psychotherapy. Constructivism in the human sciences. 2006; 11(1):150–171.

67. Chang C, Lin C. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27.

68. Gibson J, Xiao B, Georgiou PG, Narayanan SS. An audiovisual approach to learning salient behaviors in couples problem solving discussions. Proc. ICME. 2013 Jul.

Biographies

Bo Xiao (StM’08) received the Bachelor and Master degrees in Electronic Engineering Department from the Tsinghua University, Beijing, in 2007 and 2009, respectively. He is currently a PhD candidate in Electrical Engineering at the University of Southern California, Los Angeles. His current research focuses on multimodal and multimedia signal processing towards analysis and modeling of human interactive behaviors. He is broadly interested in multimedia signal processing, speech and language processing, machine learning, and human-centered computing. He is a student member of the IEEE.

Panayiotis Georgiou (StM’97-M’03-SM’10) received the B.A. and M.Eng. degrees (with Honors) from Cambridge University (Pembroke College), Cambridge, U.K., in 1996, where he was a Cambridge-Commonwealth Scholar, and the M.Sc. and Ph.D. degrees from the University of Southern California (USC), Los Angeles, in 1998 and 2002, respectively. Since 2003, he has been a member of the Signal Analysis and Interpretation Lab at USC, where he is currently an Assistant Professor. His interests span the fields of multimodal and behavioral signal processing. He has worked on and published over 100 papers in the fields of behavioral signal processing, statistical signal processing, alpha stable distributions, speech and multimodal signal processing and interfaces, speech translation, language modeling, immersive sound processing, sound source localization, and speaker identification. He has been a PI and co- PI on federally funded projects notably including the DARPA Transtac SpeechLinks and currently the NSF An Integrated Approach to Creating Enriched Speech Translation Systems and Quantitative Observational Practice in Family Studies: The case of reactivity.” He is currently an editor of EURASIP Journal on Audio, Speech, and Music Processing, Advances in Artificial Intelligence and served as a guest editor of the Computer Speech And Language, and as a member of the Speech and Language Technical Committee. He is currently the technical chair of InterSpeech 2016,

area chair for InterSpeech 2015 and has served on the organizing committees for numerous conferences. His current focus is on computational mental health, behavioral signal processing, multimodal environments, and speechto- speech translation. Papers co-authored with his students have won best paper awards for analyzing the multimodal behaviors of users in speech-to-speech translation in International Workshop on Multimedia Signal Processing (MMSP) 2006, for automatic classification of married couples behavior using audio features in Interspeech 2010, and for analysis of interpreted communication in the medical domain in International Conference on Cross-Cultural Design, HCI 2013.

Brian Baucom is an Assistant Professor in the Department of Psychology at the University of Utah. He received his doctoral degree from the University of California, Los Angeles in 2008 and completed an NIH funded postdoctoral fellow at the University of Southern California from 2010 to 2012. His research focuses on emotional and behavioral processes in romantic relationships with a particular emphasis on modeling multivariate associations between signal derived metrics of emotional expression, physiological activation, and manually annotated behavior.

Shrikanth (Shri) Narayanan (StM’88-M’95-SM’02-F’09) 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 and Psychology 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]. Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor for the *Computer Speech and Language Journal* and an Associate Editor for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, APSIPA TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING and the *Journal of the Acoustical Society of America*. He was also previously an Associate Editor of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING (2000–2004), IEEE SIGNAL PROCESSING MAGAZINE (2005–2008) and the IEEE TRANSACTIONS ON MULTIMEDIA (2008–2011). He is a recipient of a number of honors including Best

Transactions Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010–2011 and ISCA Distinguished Lecturer for 2015–2016. Papers co-authored with his students have won awards including the 2014 Ten-year Technical Impact Award from ACM ICMI and at Interspeech 2014 Cognitive Load Challenge, 2013 Social Signal Challenge, Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2013 and 2010, InterSpeech 2009 Emotion Challenge, IEEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005 and ICSLP 2002. He has published over 650 papers and has been granted sixteen U.S. patents.

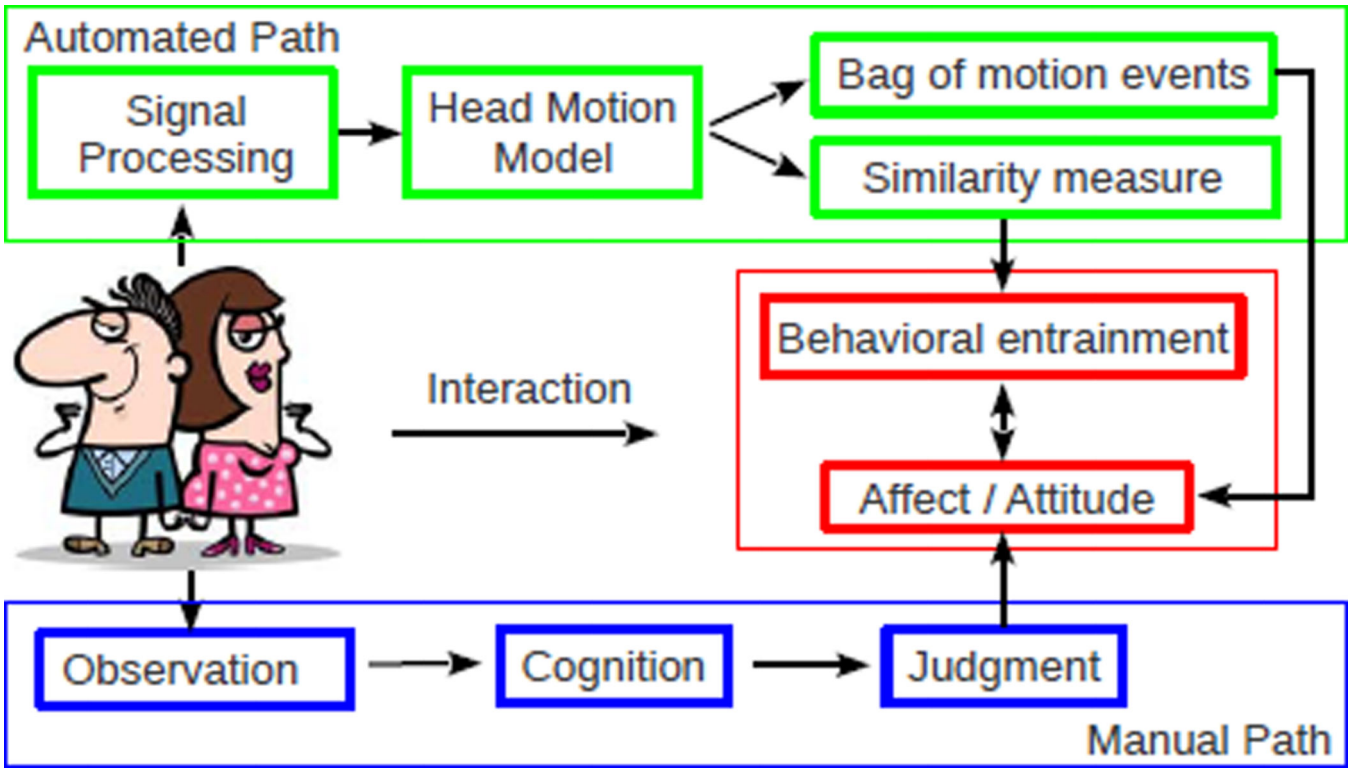

Fig. 1.

Automated path — Machine Processing, Manual Path —Human annotation, Behavioral entrainment—Head motion similarity

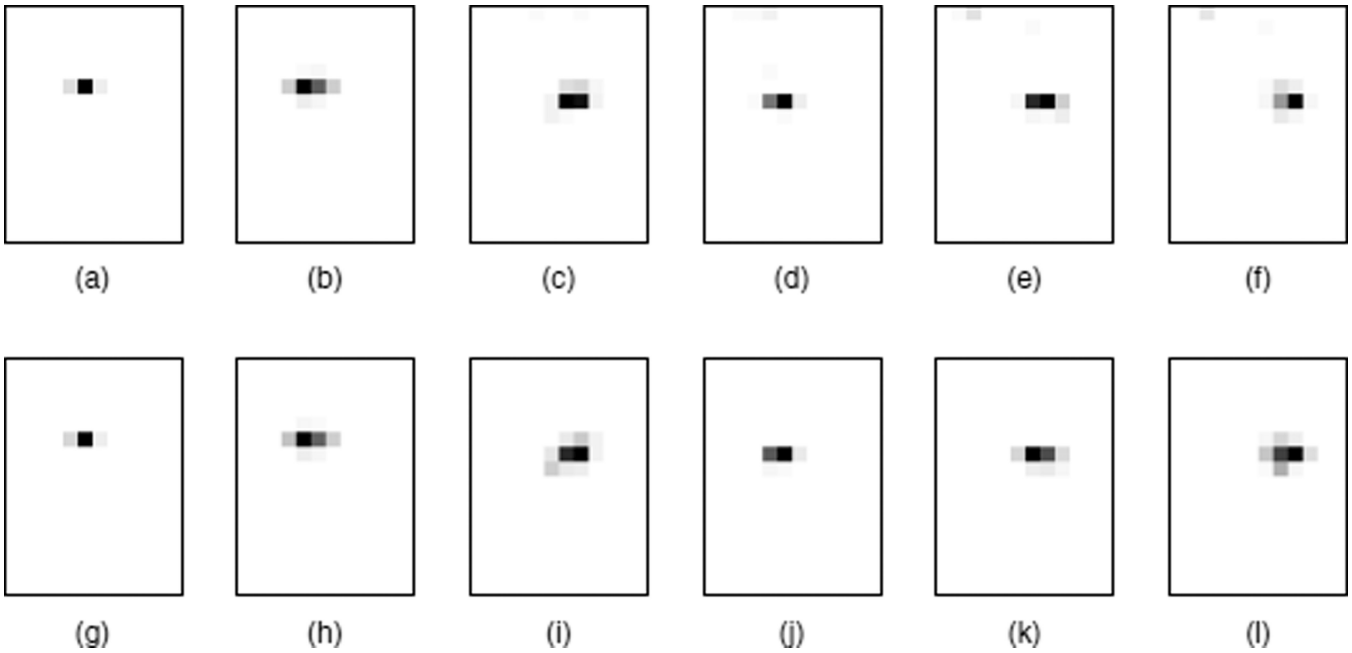

Fig. 2.

Examples of face position distribution before and after outlier removal. (a) ~ (f): before; (g) ~ (l): after; one column per session.

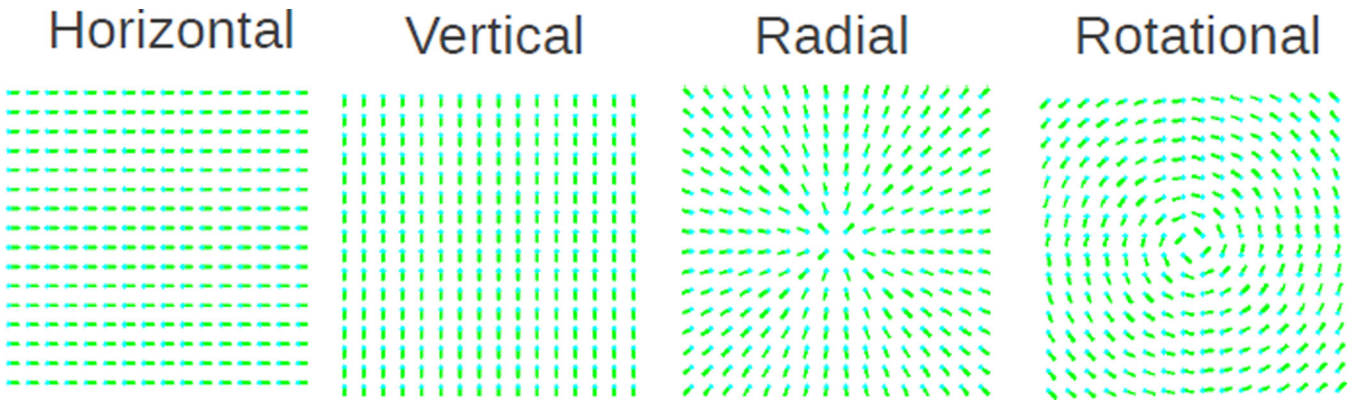

Fig. 3.
Operators on the optical flow field

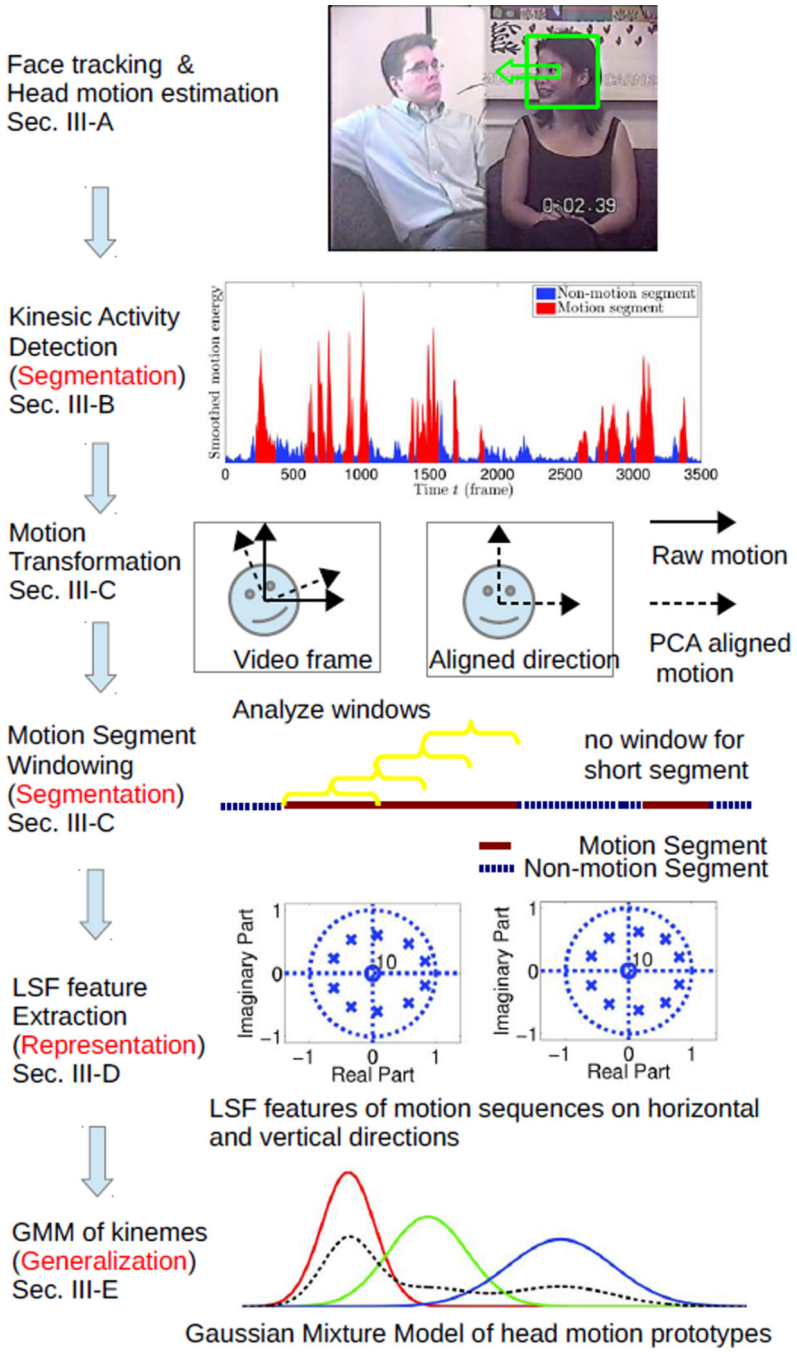

Fig. 4.
Illustration of the processing steps in Sec. III.

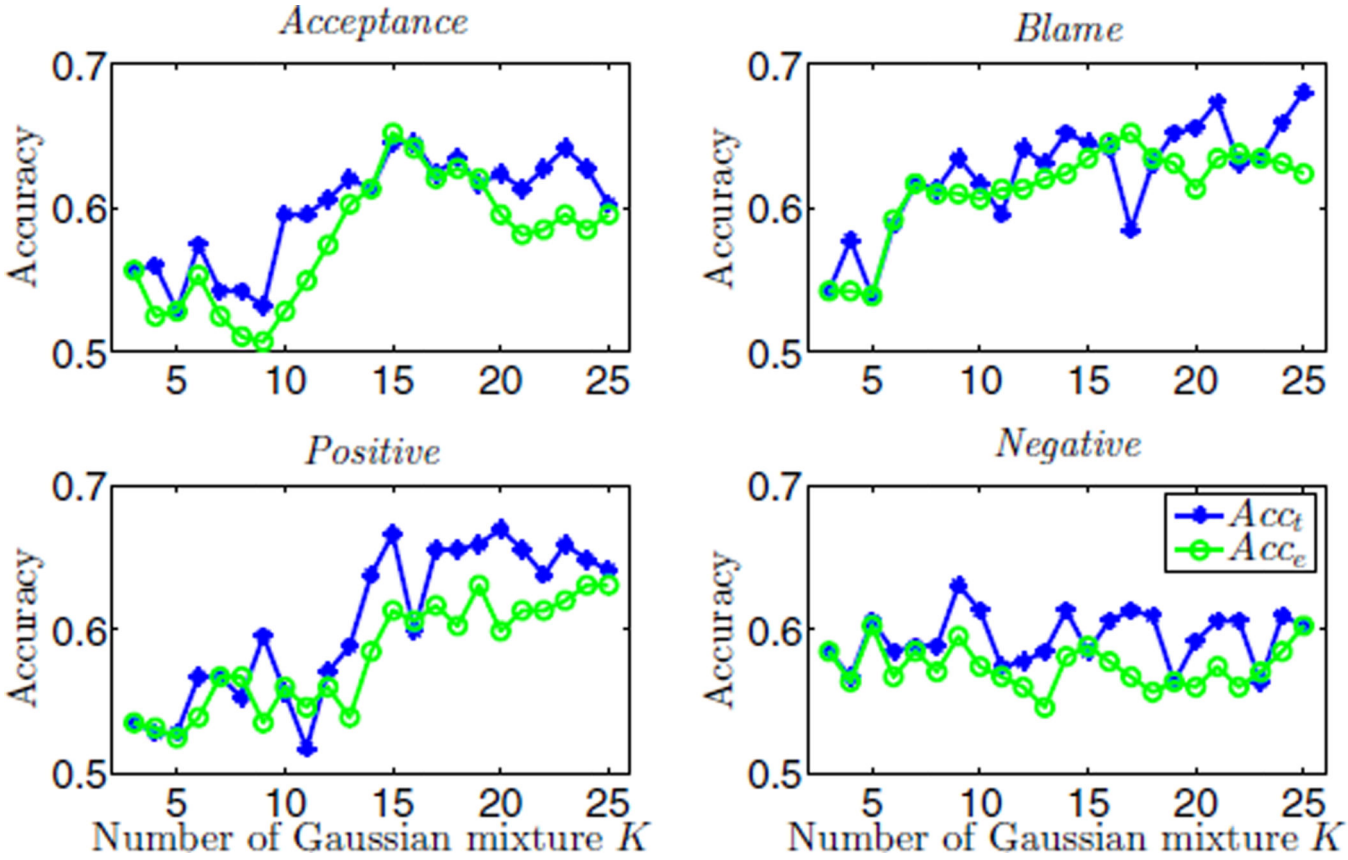

Fig. 5.

Binary classification of behavior codes using model selected by training accuracy or majority-voting of the ensemble of models

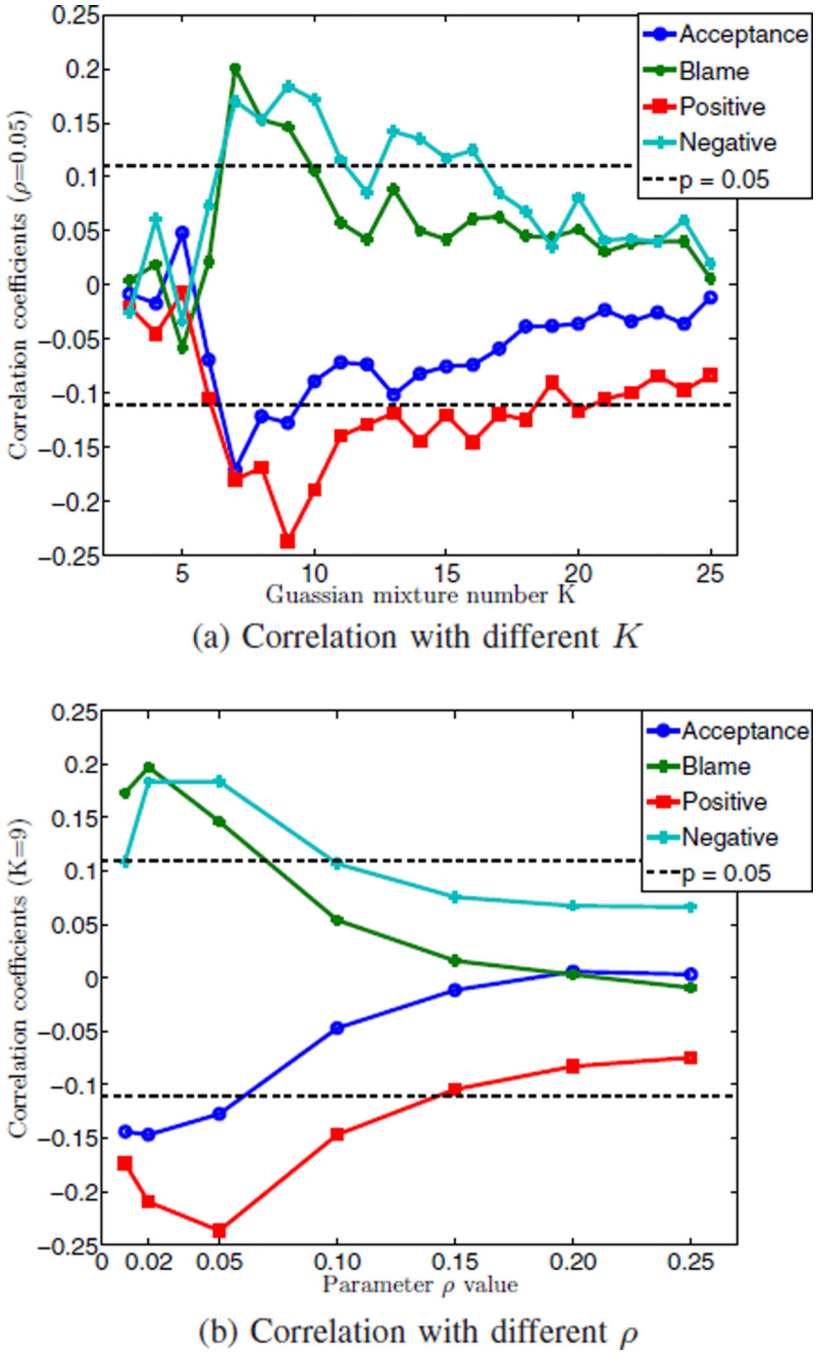

Fig. 6.
Results of hypotheses tests for the correlation between $\mathscr{R}$ and $\mathscr{Y}$

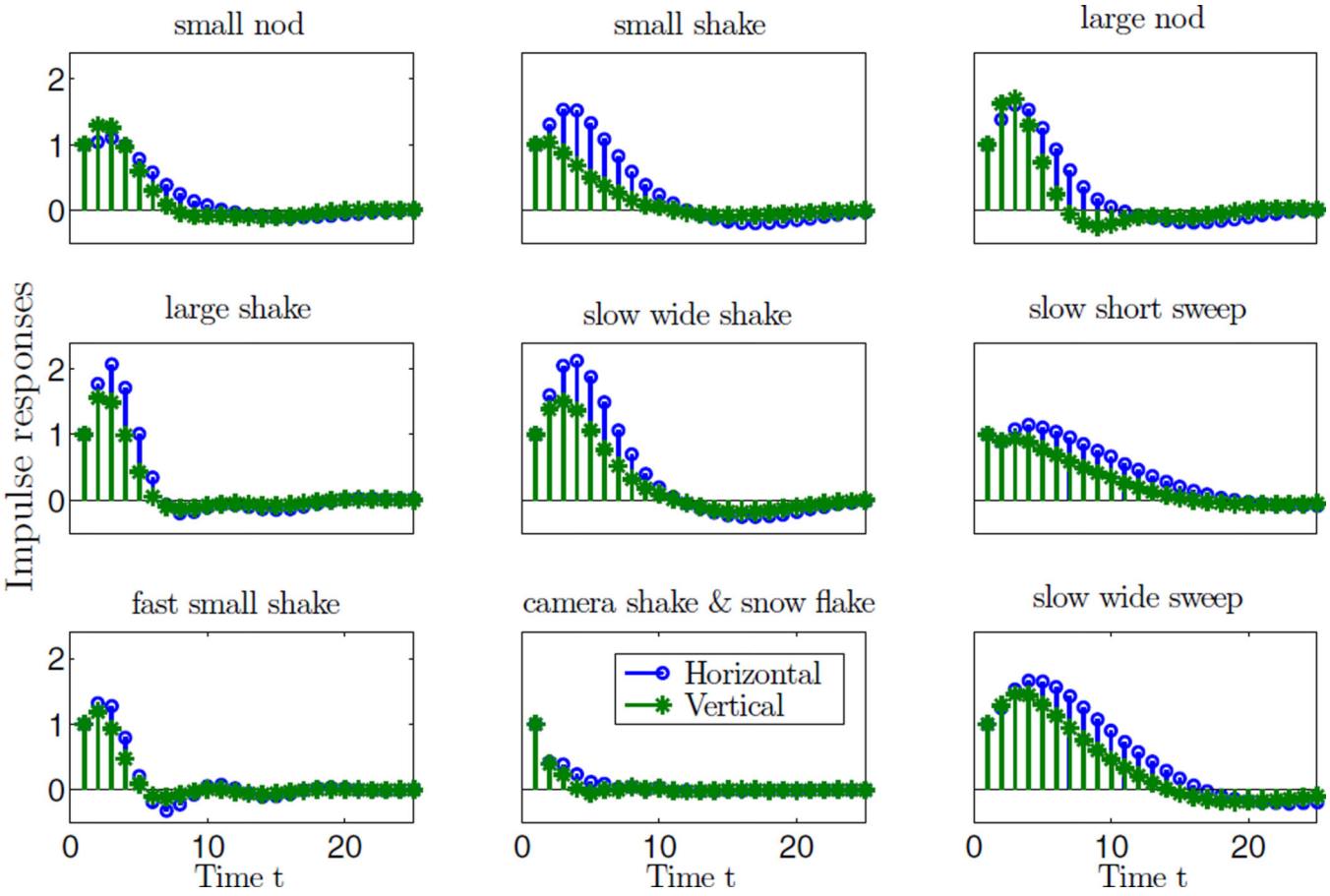

Fig. 7.

Impulse responses of the linear filters representing typical motions. Tentative descriptions: 1. small nod 2. small shake 3. large nod 4. large shake 5. slow wide shake 6. slow short sweep 7. fast small shake 8. camera shake and snow flake 9. slow wide sweep

TABLE I

Binary classification accuracies of behavior codes using 199 models selected by cross-validation

Code	*Acceptance*	*Blame*	*Positive*	*Negative*
Subjects	136	141	144	142
Sessions	282	282	282	282
Acc_{υ}	0.64	0.60	0.63	0.57
Sig.	1e–6	1e–3	1e–5	1e–2

TABLE II

Correlation between $\mathscr{R}$ and $\mathscr{Y}$ in randomized pairings

Code	*Acceptance*	*Blame*	*Positive*	*Negative*
Mean of correlation	−0.015	0.001	−0.009	0.011
Std of correlation	0.05	0.05	0.05	0.06