Multi-Party Interaction Understanding using Smart Multimodal Digital Signage

Tony Tung, *Member, IEEE*, Randy Gomez, *Member, IEEE*, Tatsuya Kawahara, *Senior Member, IEEE*, and Takashi Matsuyama, *Member, IEEE*

***Abstract*—This paper presents a novel multimodal system designed for multi-party human-machine interaction understanding. The design of human-computer interfaces for multiple users is challenging because simultaneous processing of actions and reactions have to be consistent. The proposed system consists of a large display equipped with multiple sensing devices: microphone array, HD video cameras, and depth sensors. Multiple users positioned in front of the panel freely interact using voice or gesture while looking at the displayed content, without wearing any particular devices (such as motion capture sensors or head mounted devices). Acoustic and visual information is captured and processed jointly using established and state-of-the-art techniques to obtain individual speech and gaze direction. Furthermore, a new framework is proposed to model A/V multimodal interaction between verbal and nonverbal communication events. Dynamics of audio signals obtained from speaker diarization and head poses extracted from video images are modeled using hybrid dynamical systems (HDS). We show that HDS temporal structure characteristics can be used for multimodal interaction level estimation, which is useful feedback that can help to improve multi-party communication experience. Experimental results using synthetic and real-world datasets of group communication such as poster presentations show the feasibility of the proposed multimodal system.**

***Index Terms*—human-machine system, multi-party interaction, smart digital signage, multimodal interaction dynamics**

I. Introduction

OVER the last decade smart systems for multi-party human-machine interaction have become ubiquitous in many everyday life activities. To date, multimodal Audio/Video (A/V) systems designed for human behavior and interaction analysis usually consist of multiple video cameras and microphones placed in a dedicated room, and oriented towards participants (see VACE [1], Mission Survival [2], AMI [3], IMADE [4]). Numerous hours of discussions can then be recorded for further offline analysis. Communication signals from voice and gesture are usually processed by speech and video techniques. However, these systems are still very tedious to setup and often require wearable equipment that prevent them to be used casually or in an uncontrolled environment. Moreover several steps such as video annotation still often require hours of human manual labor.

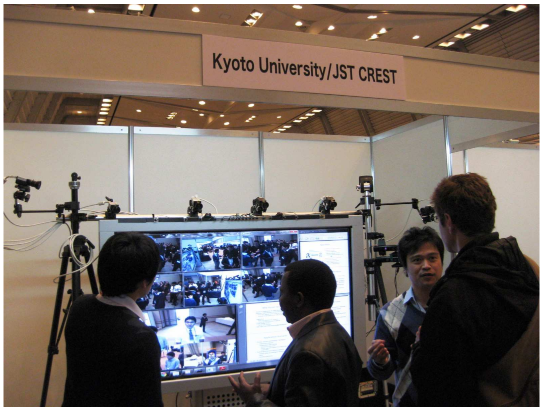


Fig. 1. Smart digital signage for group dynamics analysis showcased at the international conference IEEE ICASSP 2012. The system consists of a large display and multiple sensing devices (e.g, microphone array, HD video cameras, and depth sensors) that perform simultaneous captures. Multimodal interaction dynamics are modeled using hybrid dynamical systems, and interaction level of participants are estimated using audio-visual event statistics.

Here, a novel multimodal system that is designed for multi-party human-human interaction analysis is presented. The system is scalable and portable, and employs established and state-of-the-art techniques in speech [5], vision [6], and graphics processing (GPU) for multiple-people multimodal interaction data capture and analysis. It consists of a large display equipped with multiple sensing devices spaced on a portable structure: one microphone array, six HD video cameras, and two depth sensors (see Fig. 1). Multiple users standing in front of the panel can interact freely using voice or gesture while looking at the displayed contents, without wearing any particular device (such as motion capture sensors or head mounted devices). Using audio and visual information processing, the system detects and recognizes verbal and non-verbal communication events from multiple users (e.g., voice and head motions). The display is used to display information and trigger group discussions. Furthermore, a new framework is proposed to model A/V multimodal interaction between verbal and nonverbal communication events. Dynamics of audio signals obtained from speaker diarization and head poses extracted from video frames are modeled using hybrid dynamical systems (HDS) [7]. Thus, characteristics of state temporal structure (i.e., duration, synchronization, delays, and overlaps) can be used for interaction level analysis. In this

Manuscript received December 1, 2012; revised December 27, 2012.

T. Tung is with the Academic Center for Media Studies and the Graduate School of Informatics, Kyoto University, Kyoto, Japan (email: http://www.tonytung.org)

R. Gomez was with the Academic Center for Media Studies, Kyoto University, Kyoto, Japan

T. Kawahara is with the Academic Center for Media Studies and the Graduate School of Informatics, Kyoto University, Kyoto, Japan (email: kawahara@i.kyoto-u.ac.jp)

T. Matsuyama is with the Graduate School of Informatics, Kyoto University, Kyoto, Japan (email: tm@i.kyoto-u.ac.jp)

work, it is assumed that a multi-party interaction event such as joint attention or mutual gaze between two subjects occurs when correlated actions and reactions can be observed (e.g., assertions and head nods or turns). In the context of poster presentations, participants communicate close to each other, empirically validated using real-world video datasets.

The multi-party communication experience can therefore be improved by correcting speaker presentation, or catching users' interests by displaying specific contents. The system was designed to perform head motion tracking and speaker turn estimation, and was tested with real poster presentations as well as casual discussions. The rest of the paper is organized as follows. The next section discusses related work. Section III gives a description of the system framework. Section IV details A/V multimodal data processing. Section V presents multimodal interaction modeling using hybrid dynamical systems from verbal and nonverbal communication events (e.g., speaker turns and head motions). Section VI shows experimental results with applications to multi-party interaction. Section VII concludes with a discussion on our contributions.

II. Related Work

In group communication, humans use visual and audio cues to convey and exchange information. Video and audio data have therefore been extensively used to study human behavior in communication.

Speech is often used to detect dominant speakers based on turn-taking and behavior analysis, while nonverbal cues provide feedback to understand communication patterns and behavior subtleties and can be used as back-channels to improve communication [8], [9]. For example, several corpuses such as VACE [1], Mission Survival [2], AMI [3], IMADE [4] capture multimodal signals in multi-party conversation and interaction, while related systems explored meeting summarization applications and provide an efficient way of navigating meeting content [10], [11]. Nevertheless heavy equipment (e.g., headset or motion capture system) is also often required to obtain accurate spatial measurements, as visual information processing is usually limited by technology constraints (e.g., data size, video resolution). In the case of audio systems, multiple microphones are actually needed to perform sound-source separation. Depending on the environment condition, a good quality system may be required to achieve better sound quality performance. This involves increasing the number of microphone sensors and using high quality microphones (see details in Sect. III). Similar established sound processing systems (i.e., microphone array processing, source separation, and speech segmentation) are usually employed for robot audition [12], [13]. Thus, the system proposed next uses non-wearable sensors and established and state-of-the-art speech and computer vision technology. It addresses real-world conditions such as meeting rooms. Table I shows the number of participants involved in the experiments (#Subjects), and the evolution of sensor technology: number of cameras and resolutions (#Cams), number of microphones (#Mics), use of IR cameras and markers for motion capture (Mocap), and use of depth sensors (Depth).

TABLE I
Hardware setting of various multimodal systems.

Project	#Subjects	#Cams	#Mics	Mocap	Depth
VACE [1]	8	10VGA+	6+8	yes	no
Misson S. [2]	4	8VGA	4+1	no	no
AMI [3]	4	6VGA+	24	no	no
IMADE [4]	3	8VGA	4	yes	no
[ours]	4	6UXGA	16	no	yes

In the context of human-human interaction analysis, we are particularly interested in trigger events and behaviors. Our digital signage features a large display and can be used as a smart poster for demonstration in open space or scientific presentation at conferences (see Fig. 1). Other systems relying on a large display, such as the MIT Mood Meter (2012), usually utilize only one video camera for face detection and classification purpose: acoustic information is not used and human-human interaction is out of the scope. Commercial systems, like Sony PlayStation with EyeToy (2003), Toshiba gesture-controlled laptop Qosmio (2008), or Samsung Smart TVs (2012), use single-human gestures as remote control and do not handle interaction between multiple people.

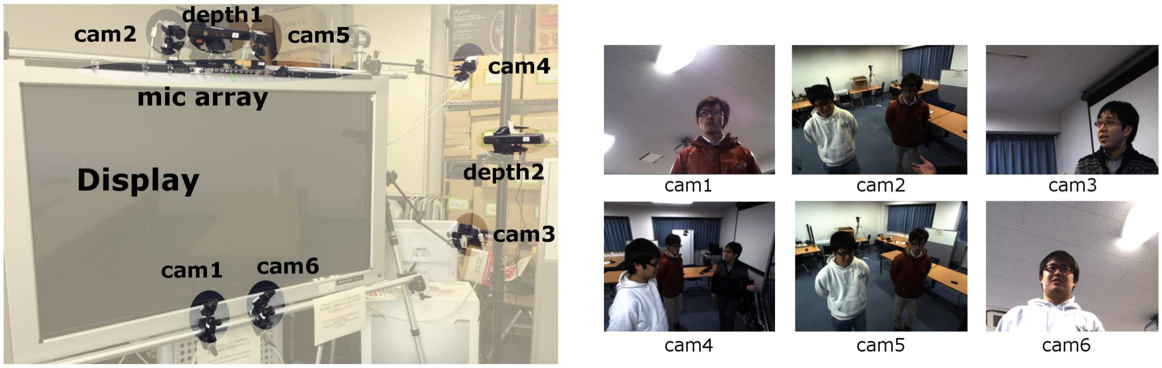
Hidden Markov Models (HMM) are very popular for speech and gesture modeling and recognition [14], [15]. However, limitations lie in the lack of flexibility for timing structure manipulation, which makes the modeling of some real-world events impractical, whereas event dynamics can be crucial to characterize human communication mechanisms. Hence, we propose to use Linear Dynamical Systems (LDS) to model communication event dynamics. In our framework, Hybrid Dynamical Systems (HDS) are used to model nonverbal behaviors which are synchronized with speech information [7].

As extensions to [16], [17], this paper contains additional details about the architecture, algorithms and models and additional experimental results using synthesized datasets for robustness assessments.

III. Smart multimodal system configuration

A. Audio system setting

To realize a hands-free system in which the participants are not constrained to using close-talking microphone, a microphone array system is employed. Multiple microphones are linearly aligned to enable multiple signal captures. Since there are multiple sources involved (i.e., interacting participants), individual signals of interests are derived through source separation. Thus, multiple signals are converted into a single signal (separated signal) belonging to either one of the participants. This system also allows the suppression of unwanted noise as it can steer its sensitivity through beamforming and ignore the signal coming from the direction of a noise source. As a result, participants can freely move away from the microphones. We note that the speech recognizer is very sensitive to the drop in SNR, and in such event, the system is vulnerable to noise contamination which degrades speech recognition performance. Thus, the use of a microphone array system supports a smooth interaction experience. We use 19 microphone sensors in total (Shure Lapel microphones) mounted on top of a 65-inch



a) Mutimodal system

b) Multiview videos

Fig. 2. a) Multimodal system setup for smart poster presentation. The system consists of one mic array, six HD video cameras, and two depth sensors to capture presenter and audience. b) Video captures of one presenter and two attendees from six views.

display (see Fig. 2a). Audio capture is performed at 16kHz sampling rate using RME HDSP audio capture device, and RME Multichannel Mic-Preamps with AD conversion.

B. *Visual system setting*

Multiple video cameras are employed to capture nonverbal communication events of multiple people standing in front of a large display. Six HD video vision cameras (Point Grey Grasshopper) are spaced on a portable structure made of poles and mounted around the display. For poster presentation, sensing devices are placed at one side of the display to capture a presenter, and at the center to capture the audience. Particularly, three cameras with 3.5mm lenses are placed on top of the display (two at the center, one on the side) to obtain a very wide field-of-view (150 degrees) as the display is very large and subjects are standing relatively close. Videos are captured in UXGA at 30fps for archiving purpose, and 3D information is recovered by multiview stereo techniques [6]. Additionally, three cameras with 12mm lenses are placed below the display (two at the center, one on the side) to capture closeup videos in SVGA of users' faces at 30fps. Furthermore, two consumer RGB-D cameras (Microsoft Kinect) are placed on top of the screen (one at the center, one on the side) to capture videos and depth maps in VGA at 30fps. The whole equipment ensures accurate and effective real-time 3D information capture of the observed scene. As only one PC with a single GPU (and three controller cards) is necessary for video capture and processing, the system is easily transportable. The system has been especially designed for capturing poster presentations, and provides high-resolution videos from multiple viewpoints that serve several purposes: visual tracking, 3D reconstruction, and image annotation. While a single Kinect camera can only capture up to two participants standing in front of a display, our system can handle up to four participants using multiple cameras.

Fig. 2 shows the current system and multiview video samples. Note that Fig. 1 shows the first prototype of the system which does not include depth sensors: all cameras are spaced on top of the display, and gaze directions are estimated from head poses only.

IV. Multimodal data processing

A. *Speech processing*

Acoustic signal processing in a multi-party system involves the processing of data captured from the microphone array and consists of two steps: 1) sound source separation, and 2) speaker diarization. An additional step of Automatic Speech Recognition (ASR) system can also be added.

1) Sound source separation: A source signal is observable to all of the microphone sensors in the microphone array system. Thus, processing is needed to effectively convert the multiple signals into a single meaningful signal (separated signal) in which noise and other unwanted sources are suppressed. Let us assume N sources (i.e., coming from participants) and M ($\geq N$) microphone sensors (in our case M=19). Let us denote $\boldsymbol{s}(\omega)$ as the input acoustic signal of N sources in frequency domain, described as $\boldsymbol{s}(\omega) = [s_1(\omega), \cdots, s_N(\omega)]^T$, where T represents the transpose operator. The received signals in vector form are denoted as $\boldsymbol{o}(\omega) = [o_1(\omega), \cdots, o_M(\omega)]^T$. The observed signal can be modeled as follows:

$$\boldsymbol{o}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega) + \boldsymbol{n}(\omega), \tag{1}$$

where $\boldsymbol{A}(\omega) \in \mathbb{C}^{M \times N}$ is the *Room Impulse Response (RIR)* in matrix form. The RIR describes the room characteristics that governs the behavior of the sound signal as it is reflected inside the room enclosure. The RIR can be measured by transmitting a series of pulse sounds to the microphones and estimating its response [18]. The background noise is denoted by $\boldsymbol{n}(\omega)$. Let us assume that $\boldsymbol{n}(\omega)$ and $\boldsymbol{s}(\omega)$ are statistically independent and uncorrelated. This assumption usually holds in real environment conditions when dealing with simple types of noise contamination. The sound sources are spatially separated using *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)*, which is a byproduct of both beamforming and blind separation [19], [20]. The separated signal is denoted as $\hat{s}^{(l)}(\omega)$.

2) Speaker diarization: The problem involving speaker diarization is based primarily on classifying the participants as either a presenter or audience. Thus, in a continuous speech

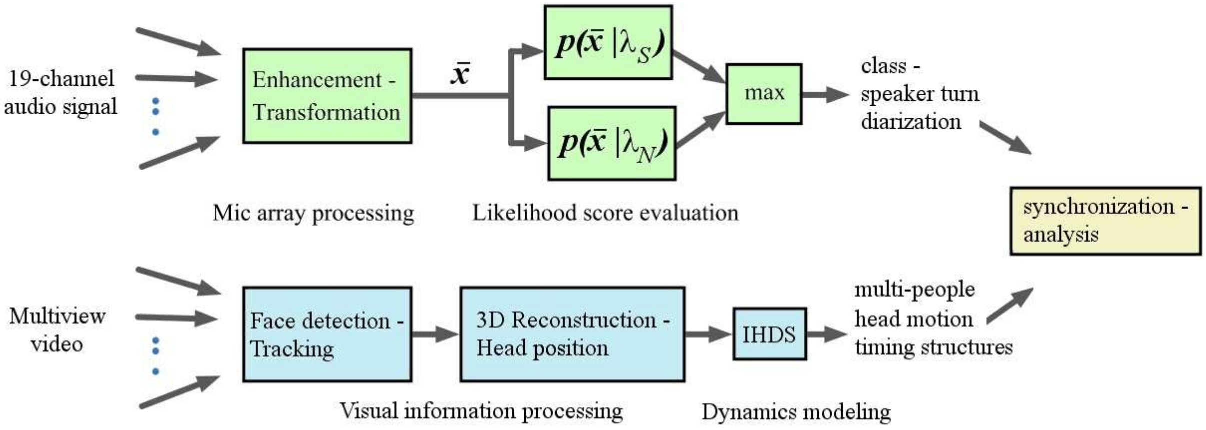

19-channel audio signal
Enhancement - Transformation
$\bar{x}$
$p(\bar{x}\,|\lambda_S)$
$p(\bar{x}\,|\lambda_N)$
max
class - speaker turn diarization
Mic array processing
Likelihood score evaluation
synchronization - analysis
Multiview video
Face detection - Tracking
3D Reconstruction - Head position
IHDS
multi-people head motion timing structures
Visual information processing
Dynamics modeling

Fig. 3. Processing scheme for multimodal interaction analysis. Top) Audio processing. Bottom) Video processing.

interaction, it is equivalent to simply identifying the presenter-audience turn. The speech signal is used as the modality in the diarization task and the performance of the system is very dependent on the speech separation quality of the two participants (i.e., audience and presenter). In real-world application, this is very difficult and separation is usually not perfect since speech itself is similar even when spoken by different people. In this work, we circumvent this problem by treating the speech-speech classification approach as speech-noise classification.

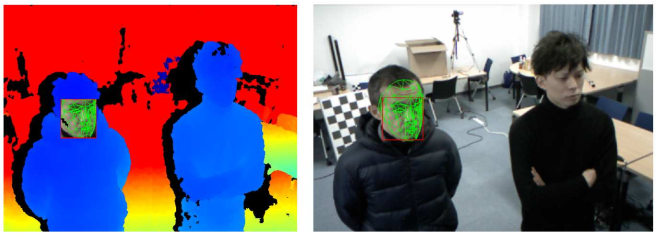
For improved classification, two classes must be distinct and fortunately, speech and noise are different types of signals. This property helps to minimize classification ambiguity. Two Gaussian Mixture Model (GMM) classifiers are designed (e.g., λ_S for speech and λ_N for noise). Depending on the size of the training dataset, the number of Gaussian mixture components are increased to improve signal discrimination. In our system, a total of 256 Gaussian mixture components are used for each model and the training of the two GMM classes is based on the Expectation-Maximization algorithm [5]. The microphone array-processed data is windowed using a 25ms frame. Then, mel-cepstrum, energy and delta energy features are extracted, and are used in the training phase.

In the actual diarization scheme shown in Fig. 3 (top), identification of the speaker turn is implemented by processing the 19-channel mic array signals resulting in $\bar{x}$ which is the separated single channel speech. Processed data contain enhanced speech (party of interest) and noise (unwanted party). Then, the likelihood score is evaluated using the enhanced speech stream against the 2 GMMs (λ_S and λ_N). Finally, the GMM that results in a higher likelihood score is selected as the corresponding class. While the conventional method only involves GMM classification (i.e., model-based only), we introduce an additional smoothing scheme after the classification procedure to improve signal discrimination between the participants during the microphone array processing at runtime. A smoothing mechanism similar to [21] is employed. The goal is to filter out erratic classification results (i.e., rapid classification changes within a small interval of time). This is based on the observation that audience-presenter switching is not likely possible within a short-time frame. Classification performance can therefore be improved experimentally by setting smoothing factors with respect to the distance to the microphone array.

3) Context and Models for ASR: In order to perform automatic speech recognition (ASR), a speaker-independent acoustic model can be trained. This approach is actually used in systems for robot audition [22], [23]. In particular, if a system is designed for poster presentation, participants are either classified as presenter or audience. Thus, at the onset of the design process, this information should be known in advance. Fortunately, this information is initially given; the presenter usually talks about the content of the presentation displayed on the digital signage. This can be used for adapting a language model for ASR. The audience is also presumed to ask questions. Because of this prior information, we can assume that the conversation dynamics between the participants are readily available. Consequently, depending on the directivity of the speaker, the acoustic model can be re-trained to $\lambda^{(l)}$ with data processed using the sound separation mechanism discussed above.

B. Multi-view video and depth processing

Visual information processing is achieved using a combination of multiple video cameras that capture the scene in high resolution (UXGA and SVGA) and depth sensors that deliver depth maps of the scene in VGA resolution. All multiview video cameras are geometrically calibrated using standard methods (i.e., with a chessboard) and synchronized by software trigger [6]. The depth sensors (IR cameras) are also calibrated with the HD video cameras. Depth data serves for human body-part identification and is performed using discriminative random forests that efficiently label different body regions [24]. Additionally, face tracking in video frames is performed using a detection-based tracker [25] relying on a cascade of weak classifiers [26], and face feature tracking and head pose (yaw, pitch, roll angles) are obtained from RGB-D data using Active Appearance Models with online learning [27] and regression models [28]. As the resolution of color cameras integrated in current consumer depth sensors is usually too poor to provide accurate gaze estimation and close range field-of-view is too limited, HD video cameras placed below the screen are used instead for accurate gaze direction estimation [29], [30]. In practice as shown in Fig. 4, regions of interest (e.g., faces, mouth and eyes) are extracted, and HD video frames are registered to depth maps.


a) Scene depth map

b) Facial feature tracking

Fig. 4. a) Depth map from frontal depth sensor. b) Face feature tracking using depth and color information (using Kinect SDK).

Compared to prior multimodal setup [16] (see Sect. II), the proposed design is able to provide more accurate head pose and potentially gaze estimation based on eye detection. Note that, as in [16], HD video cameras placed on top of the screen provide depth maps in real-time (using multiview stereo reconstruction) [6], which can be merged with data from RGB-D sensors to leverage 3D information estimation [31]. Furthermore, head and mouth positions are used in the speech processing described below for better diarization. Thus, by processing visual information the system returns temporal sequences of communication events (such as head turning and nodding) that can be correlated with temporal information from speech data.

V. Multimodal interaction dynamics

Temporal structures in speech and head motion play a crucial role in natural human communication. To date, hand-made annotations are still widely used for video analysis purpose, although extremely time consuming [32]. However, speech processing from audio data allows speaker turn diarization, and feature dynamics from visual information processing can model human communication event dynamics [7]. As shown in Fig. 3, a new model is proposed for automatic identification of repetitive behavioral patterns in multi-party multimodal interaction (e.g., such as when joint attention occurs), by exploiting timing structures (e.g., duration, synchronization, delays, and overlaps) of multimodal event dynamics.

A. Event dynamics modeling

1) Definition: A hybrid linear dynamical system (HDS) integrates both dynamical and discrete-event systems. Dynamical systems are described by differential equations and are suitable for modeling smooth and continuous physical phenomena, while discrete-event systems usually describe discontinuous changes in physical phenomena and in subjective or intellectual activities [7].

Assuming an observed *signal* (e.g., audio or video) can be discretized in atomic entities (i.e., dynamic primitives), then any (verbal or nonverbal communication) event can be modeled by: (1) a set of N linear dynamical systems (LDS) $\mathcal{D} = (D_1 \ldots D_N)$, and (2) a finite state machine (FSM) that represents states and state transitions. Let us denote a temporal sequence of an observed signal $Y = \{y(t)\}_{t=1 \ldots T}$, $y(t) \in \mathbf{R}^m$, and its hidden states $X = \{x(t)\}_{t=1 \ldots T}$, $x(t) \in \mathbf{R}^n$ belonging to a continuous state space. D_i can then be defined as:

$$\begin{cases} x(t+1) & = & F_i x(t) + g_i + v_i(t) \\ y(t) & = & H x(t) + w(t), \end{cases} \tag{2}$$

where $F_i \in \mathbf{R}^{n \times n}$ is the state transition matrix which models the dynamics of D_i, g_i is a bias vector and $H \in \mathbf{R}^{m \times n}$ is the observation matrix which maps the hidden states to the output of the system by linear projection. $v_i(t) \sim \mathcal{N}(0, Q_i)$ and $w(t) \sim \mathcal{N}(0, R)$ are process and measurement noises modeled as Gaussian distributions with zero mean and covariances Q_i and R respectively. Particularly $(F_i, H) \in \mathbb{GL}(n) \times \mathbb{ST}(m, n)$, where $\mathbb{GL}(n)$ is the group of invertible matrices of size n, and $\mathbb{ST}(m, n)$ is the Stiefel manifold. Eq. 2 is known for its ability to model complex spatio-temporal variations that possess certain temporal statistics (e.g., for dynamic textures [33], [34], human actions [35], dynamic surfaces [36], [37]). In order to control the system's state changes between two events, an FSM having a discrete set of states $\mathcal{S} = \{s_i\}_{i=1 \ldots N}$ is coupled to $\mathcal{D}$, where each s_i corresponds to an LDS D_i. The number N of LDS and their parameters $\{\theta\}$ can be estimated by clustering of LDS and optimization of $\{\theta\}$ by Expectation-Maximization [38] [7].

2) Dynamical system model representation: Interval-based representation of Hybrid Dynamical Systems (IHDS) allows us to describe event timing structures of model states (i.e., duration, synchronization, delays, and overlaps). They can also be used for event classification or recognition [7]. Let us denote $I_k =< s_i, \tau_j >$ an interval identified by a state (or mode) $s_i \in \mathcal{S}$ and a duration $\tau_j = e_k - b_k$, where b_k and e_k are the starting and ending time of I_k respectively. Complex human behaviors (i.e., verbal and nonverbal communication events) are represented using IHDS (see Fig. 3 (bottom)), similarly to a musical score where $\{I_k\}$ are notes and N is the scale. (Hence, as s_i and thus D_i is activated, a sequence of continuous states can be generated from $\{x(t)\}$ and mapped to the output observation space as $\{y(t)\}$.)

B. Multimodal interaction level modeling

Let us define an interaction event as an action-reaction pair. Particularly, the *interaction level* between multimodal signals can then be defined by the number of occurrences of synchronized events that happen within a delay (i.e., reaction time), and can characterize reactivity. Synchronized events can be identified by computing temporal differences between the beginning and ending of each interval. Hence, signal synchronization Z of two signals Y_k and $Y_{k'}$ can then be estimated by identifying all overlapping intervals (i.e., synchronized events) in the signal $\mathcal{I} = \{(I_k, I_{k'}) : [b_k, e_k] \cap [b_{k'}, e_{k'}] \neq \emptyset\}$, and by considering the following distribution:

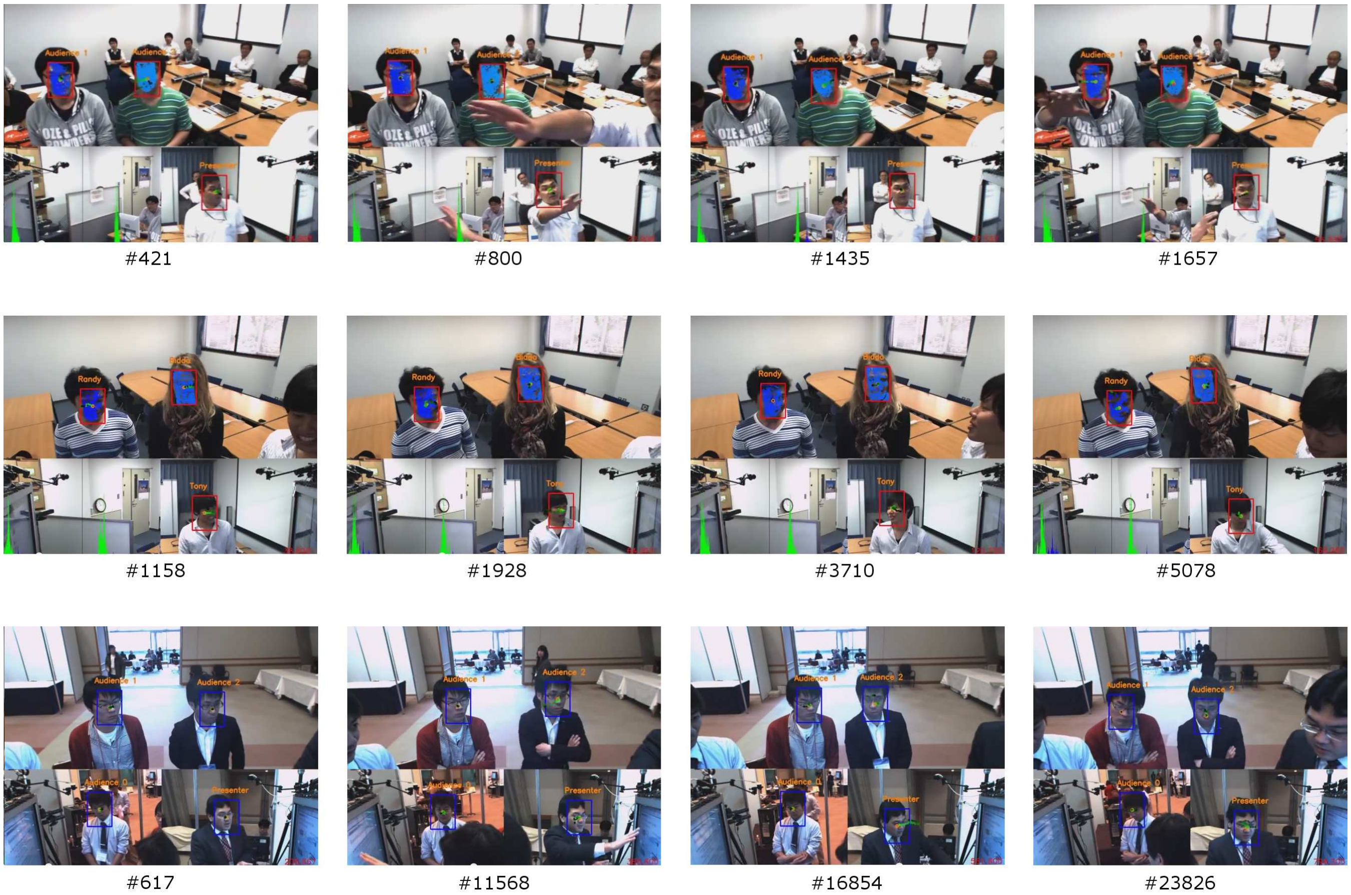


Fig. 5. Examples of poster presentations and group discussions captured by our multi-view video system in various environments. In each image, top view shows audience, and bottom-right view shows presenter. Frame number is given below each image (cameras are synchronized). Faces and face features are tracked across sequences to determine head motions, while speech is recorded. (Top row) Poster presentation held in a meeting room (one presenter, two in the audience). Bottom-left view is left for statistical representation of pixel depth (i.e., blue dots obtained from online multiview stereo reconstruction) which is used for tracking of each audience member's face region. (Middle row) Casual discussion held in a meeting room (one presenter, two in the audience). (Bottom row) Poster presentation held in a conference hall with high ceiling and dim lighting (one presenter, three in the audience). Bottom-left view shows the third audience member.

$$Z(Y_k, Y_{k'}) = Pr(\{b_k - b'_k = \Delta_b, e_k - e_{k'} = \Delta_e\}_{\mathcal{I}} | \{[b_k, e_k] \cap [b_{k'}, e_{k'}] \neq \emptyset\}_{\mathcal{I}}]). \tag{3}$$

The distribution can be modeled as a 2D Gaussian centered in $Z_0 = \frac{\sum \Delta(I_k, I_{k'})}{N_{kk'}}$, where $N_{kk'}$ is the number of overlapping intervals in $\mathcal{I}$ and $\Delta(I_k, I_{k'}) = ((b_k - b_{k'}), (e_k - e_{k'}))$ is the temporal difference between I_k and $I_{k'}$. Z contains information about reactivity with respect to reaction time (which is particularly meaningful when $|b_k - b_{k'}| < 1s$). Note that, if $\{(b_k - b_{k'}) \rightarrow 0\}$ and $\{(e_k - e_{k'}) \rightarrow 0\}$, then all pairs of overlapping intervals are synchronized.

VI. Experimental results

A. Datasets

To assess the collective performance of our framework, the system was tested in real-world situations such as in conference hall and meeting rooms (see Fig. 1,5). Our experiments focus on bipartite interactions in group discussions (i.e., between one presenter and one audience's member). Real poster presentations as well as casual discussions between 3-4 people were held to evaluate the interaction level of subjects using the proposed smart digital signage. The number of participants was limited to 4 as it was observed that close distance communication is suitable to trigger numerous nonverbal interactions. More than 20 poster presentations of around 15 minutes each involving different speakers, audiences, and poster contents were actually captured for our experiments.

B. System performance

The proposed system is installed in a small environment (2m*1.5m) and consists of 6 video cameras with UXGA resolution (1600*1200 pixels), which allows us to perform more accurate face detection and face feature tracking (see Fig. 5). Despite fairly cluttered backgrounds and various illumination conditions, the system could effectively achieve multimodal data capture. Audio, multiview videos, and depth maps are recorded simultaneously while participant detection and face feature tracking are performed in real-time. In the current implementation, multimodal interaction modeling is computed offline. The technical challenges mainly deal with storage management (e.g., 15min of continuous capture from a single camera in HD raw format at 30fps requires 20GB), and simultaneous recording and writing. As the portable system uses only one PC to perform multiple video recording, several SSD drives are used for online capture, and USB3.0 devices for data transfer to storage devices, BOOST C++ library for multithreading (for simultaneous video frame capture and writing), and GPU (NVIDIA) computing for online visual processing (e.g., face feature tracking).

C. Visual information dynamics model evaluations

First, the performance of each implemented part was evaluated individually. Face detection and face feature tracking can be performed robustly using techniques described in Sect. IV, and no major problem was encountered with real-world data. Average face feature position estimation error is below 5 degrees. Our sole recommendation is to place the system in a location with adequate lighting for better archiving quality (and as Kinect depth sensors do not perform well in direct sunlight).

To evaluate our multimodal interaction dynamics model employing Hybrid Dynamical Systems (HDS), experiments were performed on synthesized datasets that possess temporal statistics. In our framework, face features tracked across time are extracted from color and depth image information. In order to simulate real data, several 3D surfaces were created which undergo various deformations and noise and whose points can be tracked across time. The observed deformations consist of noisy sinusoidal signals $\{y\}$, where $y = \alpha \sin(\beta \pi x)$, $\alpha = 6, 4, 2, 8$ and $\beta = 10, 2, 7, 5$ respectively, with an additional random noise on x and y ($< 5\%$). Hence, time series can be extracted from surface point features, modeled using HDS and then reconstructed to evaluate the accuracy of the modeling. In our experiments, to account for spatial noise, surfaces are sampled into patches (consisting of 4 to 6 vertices) and multivariate observations are extracted locally (per vertex in each patch) from the 3D surface models. The shape index [39] is used to capture continuous local deformations, while returning values in $[0, 1]$ (see Fig. 6 (Top)). Furthermore, the same experiments were performed on 3D surfaces from real datasets obtained from multiview 3D reconstruction (as described in Sect. III-B) [40], and patches were extracted on body part regions (e.g., head, limbs) for evaluations. Note that HDS modeling is sensitive to sampling and clustering parameters, and also the number N of HDS states. However these parameters can be efficiently derived by the Expectation-Maximization algorithm (see [7], [38]). In Fig. 6, we show results on synthesized and real datasets using the same parameter set (with $N = 6$ states) that was optimized using the real datasets for the sake of consistency, by minimizing the following model fitting error which approximates the overall log-likelihood score:

$$Err(N) = \|Y - Y^{\text{rec}}(N)\|_2 = \sqrt{\sum_{t=1}^{T}(y_t - y_t^{\text{rec}}(N))^2}, \quad (4)$$

where Y is the original observed signal, and $Y^{\text{rec}}(N)$ is a reconstructed signal using N dynamical systems. The optimal number of dynamical systems is then determined by extracting the value where the error difference between consecutive steps is maximal:

$$N = \arg\max_{N_k \in \mathbb{N}^+} |Err(N_k - 1) - Err(N_k)|. \quad (5)$$

With both synthesized and real datasets, the interval-based representation shows internal patterns that repeat several times, which correspond to cycles of transitions between linear dynamical systems, and $Err(6) < 6\%$ on average after normalization. a) and d) show examples of two multivariate observations from synthesized data where $(\alpha, \beta) = (6, 10)$ and $(8, 5)$ respectively, b) and e) show HDS modeling using $N = 6$ states, and c) and f) show reconstructed signals from the HDS models, g) and j) show examples of observed signals from two point patches of real data from leg and head regions respectively, h) and k) show HDS modeling using $N = 6$ states, and i) and l) show reconstructed signals from the HDS models. Thus, the proposed implementation can effectively model visual event dynamics.

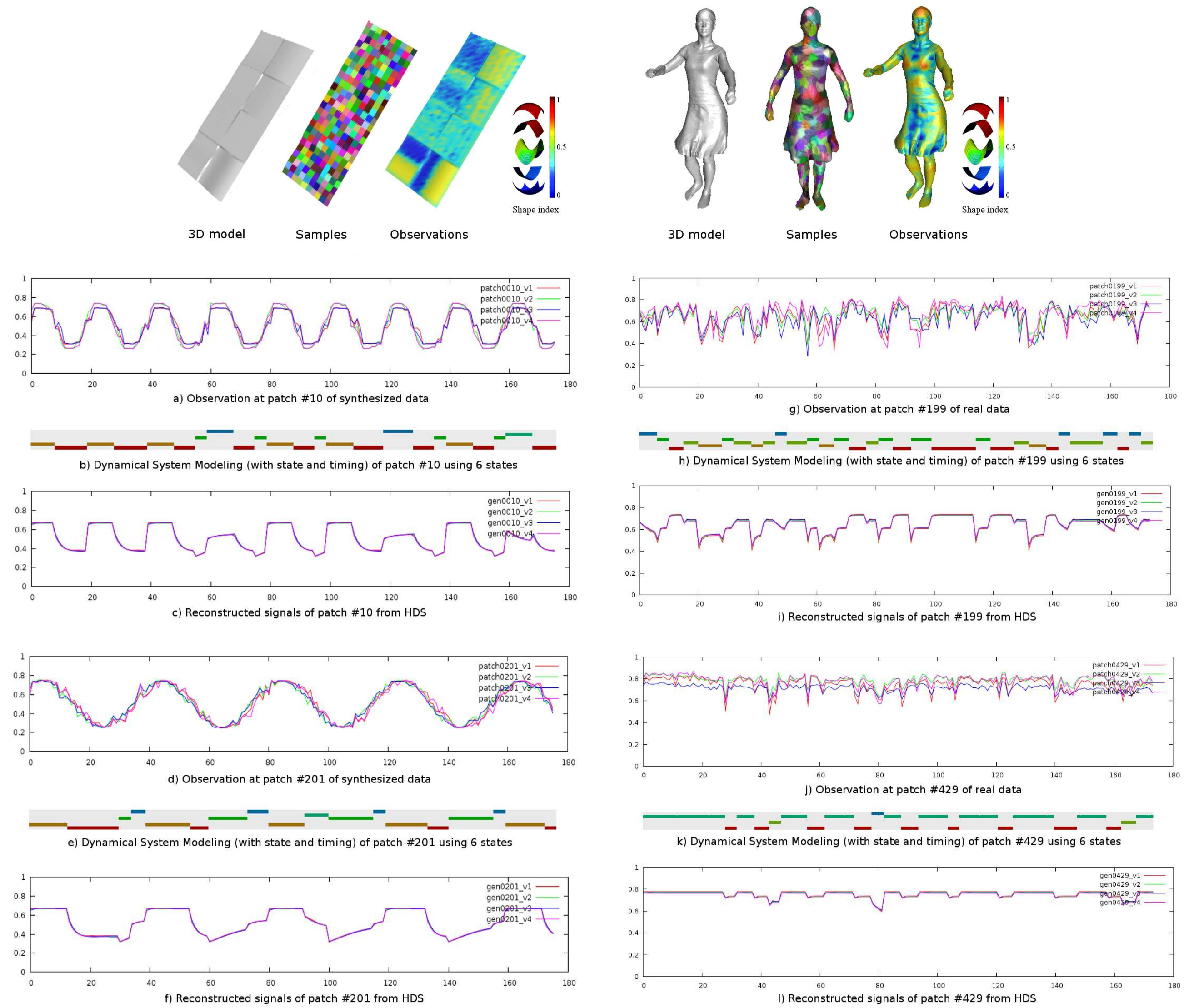
D. Speaker diarization results using real data

The objective of speaker diarization is to track speaker turns during the whole poster presentation. We note that the recorded speech signal contains both the audience and the presenter respectively. Thus, the corresponding speech segments belonging to either the presenter or the audience is tracked. The performance of the proposed classification depends primarily on the following processes:

- The quality of the source separation through the microphone array processing.
- The quality of noise-rendering to the unwanted signal. Note that the speech other than the person of interest is transformed to noise for improved discrimination.
- The acoustic model training used for the GMM classification.

In Fig. 7, we show the results of the classification performance using real recording when the audience is standing at distances 1.0m, 1.5m and 2.0m, respectively with an angle of -30 degrees relative to the normal axis of the display. The presenter position is fixed at approximately 0.5m and +30 degrees relative to the normal axis of the display. We record a total of 100 samples of mixed stream of speech signal (presenter and audience) for each distance location (i.e., 1.0m, 1.5m and 2.0m). Each recording is of 3-minute duration, and smoothing factors at various distances from the microphone array are obtained experimentally: 0.10s at 1.0m, 0.20s at 1.5m and 0.35s at 2.0m (see Sect. IV-A2 for details). The classification results (averaged over total samples) in Fig. 7 show the performance of identifying both the presenter and audience separately using our classification scheme with noise-rendering, in comparison to a conventional method that employs only GMM classification (model-based only).

Moreover we conducted statistical significance test based on t-statistics for equal sample sizes with unequal variances [41]. To perform this, an additional set of 100 recording samples is collected. T-tests showed that the improvement due to our



a) Observation at patch #10 of synthesized data

b) Dynamical System Modeling (with state and timing) of patch #10 using 6 states

c) Reconstructed signals of patch #10 from HDS

d) Observation at patch #201 of synthesized data

e) Dynamical System Modeling (with state and timing) of patch #201 using 6 states

f) Reconstructed signals of patch #201 from HDS

g) Observation at patch #199 of real data

h) Dynamical System Modeling (with state and timing) of patch #199 using 6 states

i) Reconstructed signals of patch #199 from HDS

j) Observation at patch #429 of real data

k) Dynamical System Modeling (with state and timing) of patch #429 using 6 states

l) Reconstructed signals of patch #429 from HDS

Fig. 6. Examples of local multivariate observations, interval-based representation of HDS models, and signal reconstructions from HDS for synthesized and real objects (Left and Right respectively). a) and d) are observations from simulated noisy data, while g) and j) are observations extracted from real human data. b), e), h) and k) show interval-based representations of HDS modeling, respectively. c), f), i) and l) show reconstructed signals, respectively.

proposed classification was significant: $t(198) = -2.13, p = 0.034$.

E. Interaction level analysis

Multi-party interaction events (that include joint attention and mutual gaze) are located and estimated from synchronized overlapping verbal and nonverbal communication events between participants (i.e., when multimodal interaction level is high). In order to provide evaluations against ground-truth measurements, the sequences were annotated manually [32]. The following results were obtained on two representative sequences of the datasets. The sequence samples shown in Fig. 8 contain 2000 and 2500 frames respectively. a) show the results for a technical poster presentation involving one presenter and two in the audience, and b) a casual discussion between three subjects commenting on photos displayed on the digital signage.

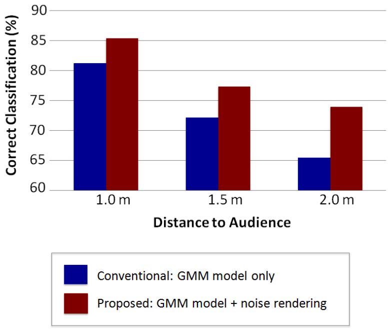


Fig. 7. Classification results using 100 test recording samples of real speech data at various distances. The conventional method only involves GMM classification (i.e., model-based only). On the other hand, we incorporate noise-rendering to improve signal discrimination between the participants.

Head motion dynamics were modeled using HDS, where time series are obtained from head pose angles (see plots in pixels: x is yaw, y is pitch). Interval-based representations of HDS model states (IHDS) are computed with $N = 4$ states. Here, we observed that state changes correspond to specific head motions (e.g., turning to the display or audience and nodding). The number of head movements of the participants as well as their reactions to visual and audio stimuli are then counted for comparison. Results are reported in Tab. II and Tab. III. Presenter in a) and Subject 1 in b) made numerous head movements towards the screen, and towards other participants. In a) Audience 2 produced much more nonverbal communication events than Audience 1, while in b) Subject 2 and Subject 3 behaved similarly.

TABLE II
STATISTICS FOR MULTI-PARTY INTERACTION IN SEQUENCE A).

	Presenter	Audience 1	Audience 2
Head movements (/min)	26.5	32.0	19.5
Speech turns (%)	62.5	3.1	9.4
Visual reactions (/min)	-	33	46
Audio reactions (/min)	-	5	8

TABLE III
STATISTICS FOR MULTI-PARTY INTERACTION IN SEQUENCE B).

	Subject 1	Subject 2	Subject 3
Head movements (/min)	32.3	25.3	27.3
Speech turns (%)	73.5	14.2	12.3
Visual reactions (/min)	-	83	87
Audio reactions (/min)	-	13	11

In addition, Fig. 8 (right) shows the timing structure synchronization distributions for all pairs of overlapping states obtained from HDS modeling as described by Eq. 3. Each point of a plot represents the synchronization disparity of a pair of overlapping states. Hence a point at $(0, 0)$ stands for total synchronization of a pair of states, meaning that two events (head motions) performed by two subjects are performed simultaneously, and with the same duration. A point located in the upper-right quarter stands for $\Delta b > 0$ and $\Delta e > 0$. Red circles locate distribution centers and represent the synchronization disparity average. The temporal differences have a maximum $|\Delta b|$ and $|\Delta e|$ of 60f (i.e., 4s). In the case of a), disparity is lower between head movement dynamics of Presenter and Audience 2, than between Presenter and Audience 1 as more points are located closer to the center. In the case of b), disparity is similar between head movement dynamics of Subject 1 and Subject 2, and between Subject 1 and Subject 3 (both distributions are similar). Timing structure synchronization distribution details for each combination of pair of overlapping states are given in Fig. 9.

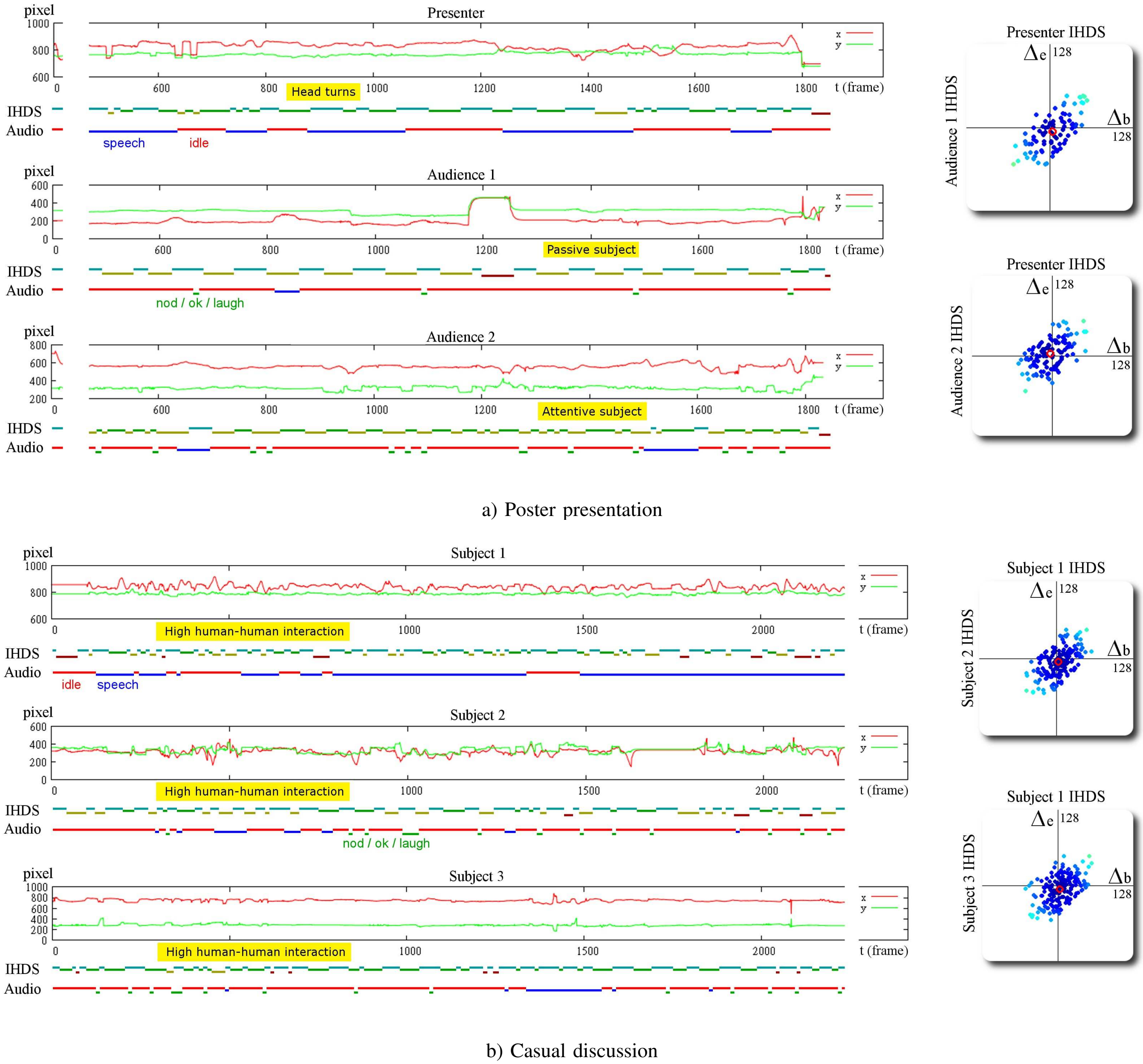
From these observations, the interaction levels between participants can be measured (as defined in Sect. V): in the poster presentation, Presenter interacted much more with Audience 2 than with Audience 1, while in the Casual group discussion, both Subject 2 and Subject 3 interacted equally with Subject 1. Moreover, multimodal interaction levels between speech and head movements are also modeled. Speaker turns obtained by diarization (see Sect. V) allow us to evaluate the speech activity of each participant. In a) Presenter speaks more, while Audience 2 speaks more than Audience 1, and in b), Subject 1 speaks the most, while Subject 2 and Subject 3 perform similarly. Global statistics for multi-party multimodal interaction analysis obtained from state change occurrences are reported in Tab. II and Tab. III. Naturally, the accuracy of these measures depends on the ability of the vision system to track faces and estimate orientations, and the ability of the speech system to perform speaker diarization. In our experiments, we globally obtained very accurate estimations compared to hand-made annotations of audio and visual events ($< 5\%$ errors). Let us mention that the face tracking of Audience 1 in a) was lost around frame 1200 during the processing due to an implementation issue. Nevertheless the unexpected tracking behavior is successfully identified as a separate state by the HDS model.

Furthermore, multimodal interaction levels are computed between each participants in order to determine the reactivity to visual and audio stimuli. Figure 10 presents interaction levels between the main speaker and each participant. The graphs represent the total number of reactions synchronized with stimuli with respect to the number of frames (from the beginning of the sequences), with a video frame rate of 15fps. In a) and b), the figure shows: (Left) head reactions in response to audio stimuli for all participants (main speaker included), and (Right) head reactions of participants other than the main speaker in response to visual stimuli from him. In a), it can be seen again that Audience 2 has much more reactions than Audience 1 for both audio and visual stimuli. In b), the number of reactions are similar, showing equal interaction level between Subject 2 and Subject 3. Note that more reactions are found with the visual stimuli. As human reaction time to audio and visual stimuli is usually below 1s (15 frames), the level of attention of each participant can be derived by the behavior of the curves near the origin. Interestingly it can be observed that reaction times of Audience 2 and Subject 2 are very good, which is also confirmed by checking the videos. As a result, multimodal interaction dynamics modeling allows us to locate and estimate multi-party interaction events (e.g., that include joint attention and mutual gaze) from synchronized overlapping verbal and non-verbal communication events between participants of poster presentation or casual group discussion.

VII. CONCLUSION

This paper presents a novel multimodal system that is designed for multi-party human-human interaction analysis. Human speech communication is intrinsically bi-directional and duplex, and feedback behaviors play an important role in smooth communication. Feedback behaviors of an audience are particularly important cues in analyzing presentation-style conversations.

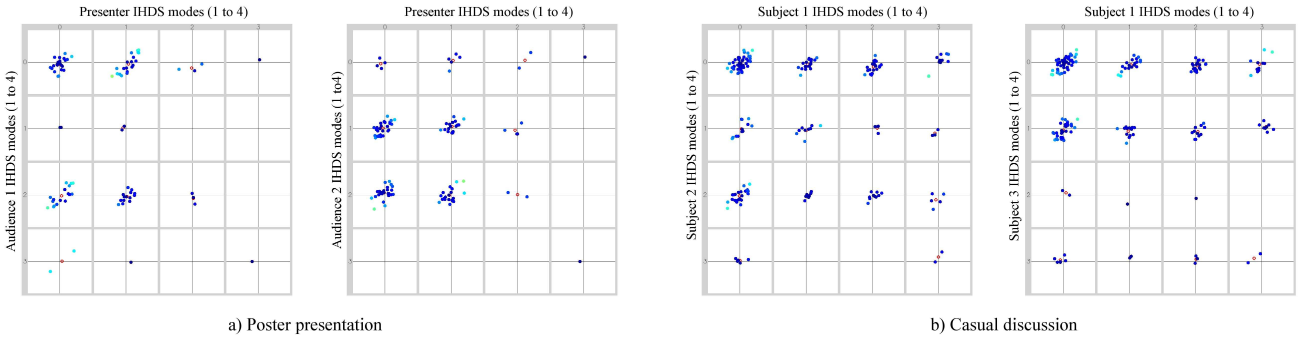
In this work, the proposed system which consists of a large display equipped with multiple sensing devices has been tested in real-world situations such as poster presentations,



a) Poster presentation

b) Casual discussion

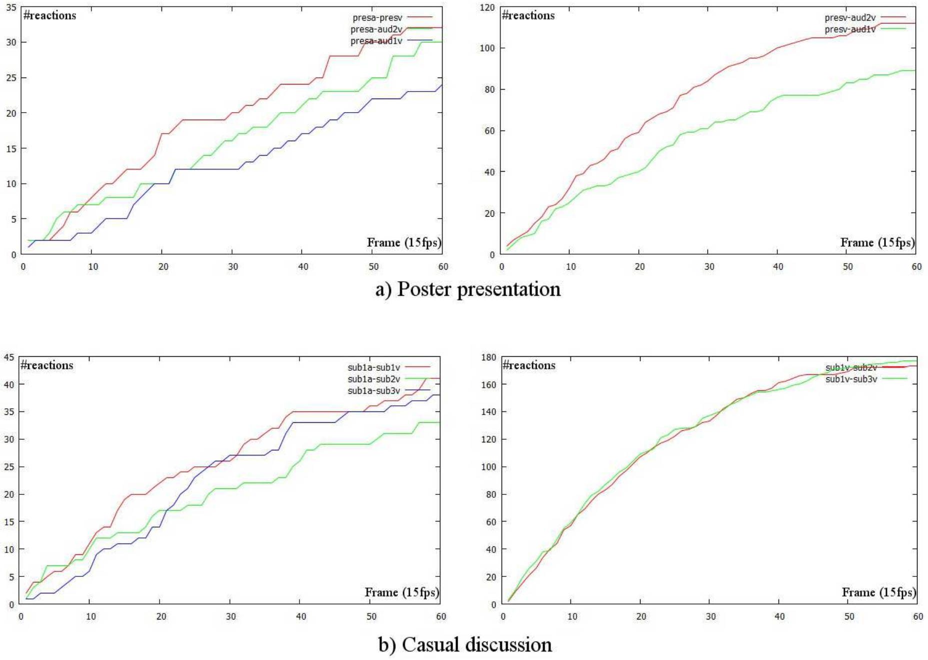
Fig. 8. Multi-party multimodal interaction dynamics modeling for: a) Poster presentation and b) Casual discussion. From the top: head position (x, y) in pixels, IHDS modeling with 4 modes, and speaker diarization (red: idle, blue: speech, green: nod/ok/laugh). Right: IHDS synchronization distributions.



a) Poster presentation

b) Casual discussion

Fig. 9. Timing structure synchronization modeled using multivariate normal distributions. Each point of a plot represents the synchronization disparity of a pair of overlapping states $(\Delta b, \Delta e)$. Colors (light to dark blue) represent distances to center (red). In a) Poster presentation, synchronized communication events are more frequent between Presenter and Audience 2 than with Audience 1. In b) Casual discussion, statistics of Subject 2 and Subject 3 are similar.



a) Poster presentation

b) Casual discussion

Fig. 10. Multimodal interaction level for a) Poster presentation, and b) Casual discussion. The graphs represent the total number of reactions synchronized with stimuli with respect to the number of frames.

in which a researcher makes an academic presentation to a couple of persons using a digital poster. Poster sessions have become a norm in many academic conventions because of the interactive characteristics. Audio and visual information is captured and processed jointly using established and state-of-the-art techniques to obtain individual speech and gaze direction, while multiple users positioned in front of the panel freely interact using voice or gesture, looking at the displayed contents. In addition, a new framework is proposed to model A/V multimodal interaction between verbal and nonverbal communication events using hybrid dynamical systems. In particular, we show that visual information dynamics can be used to detect nonverbal communication events that are synchronized with verbal communication events. As a result, multimodal interaction dynamics modeling allows us to estimate users' interaction level. Speaker presentation or displayed information can then be adapted for better communication.

For future research, we are investigating a multimodal system in wider area such as meeting rooms, where more participants and interactions would be involved, and full 3D data could be considered [42]. We are also working on smaller scale systems where a compromise on accuracy has to be done with low resolution equipments.

Acknowledgment

This work was supported by the JST-CREST project Creation of Human-Harmonized Information Technology for Convivial Society. The authors would like to thank all the project members for valuable inputs, and the numerous participants who helped validating the system.

References

[1] L. Chen, R. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang, "Vace multimodal meeting corpus," *Machine Learning for Multimodal Interaction, LNCS Springer, S. Renals and S. Bengio Eds.*, 2006.

[2] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of concensus decision making meetings," *Language Resources and Evaluation*, pp. 409–429, 2007.

[3] M. Poel, R. Poppe, and A. Nijholt, "Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction," *IEEE Int'l Conf. Automatic Face and Gesture Recognition (FG)*, 2008.

[4] Y. Sumi, M. Yano, and T. Nishida, "Analysis environment of conversational structure with nonverbal multimodal data," *ICMI-MLMI*, 2010.

[5] R. Gomez and T. Kawahara, "Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood," *IEEE Trans. Audio, Speech and Language Processing*, 2010.

[6] T. Matsuyama, S. Nobuhara, T. Takai, and T. Tung, "3d video and its applications," *Springer*, 2012.

[7] H. Kawashima and T. Matsuyama, "Interval-based modeling of human communication dynamics via hybrid dynamical systems," *NIPS Workshop on Modeling Human Communication Dynamics*, 2010.

[8] S. White, "Backchannels across cultures: A study of americans and japanese," *Language in Society*, vol. 18, pp. 59–76, 1989.

[9] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," *Conf. Interspeech*, 2012.

[10] M. A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 1997.

[11] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," *Conf. Interspeech*, 2005.

[12] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," *National Conf. Artificial Intelligence (AAAI)*, pp. 832–839, 2000.

[13] "Hark website http://winnie.kuis.kyoto-u.ac.jp/hark."

[14] L. R. Rabiner, "A tutorial on hidden markow models and selected applications in speech recognition," *IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[15] C.-D. Liu, Y.-N. Chung, and P.-C. Chung, “An interaction-embedded hmm framework for human behavior understanding: With nursing environments as examples,” *IEEE Trans. Information Technology in Biomedecine*, vol. 14, no. 5, pp. 1236 – 1246, 2010.

[16] T. Tung, R. Gomez, T. Kawahara, and T. Matsuyama, “Group dynamics and multimodal interaction modeling using a smart digital signage,” *European Conf. Computer Vision (ECCV), Ws/Demos, Lecture Notes in Computer Sciences (LNCS), Springer*, vol. 7583, pp. 362–371, 2012.

[17] ——, “Multi-party human-machine interaction using a smart multimodal digital signage,” *Int’l Conf. Human-Computer Interaction (HCI), Lecture Notes in Computer Sciences (LNCS), Springer*, 2013.

[18] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, “An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses,” *Journ. Acoustical Society of America*, no. 2, pp. 1119–1123, 1995.

[19] H. Sawada, R. Mukai, S. Araki, and S. Makino, “Polar coordinate based nonlinear function for frequency-domain blind source separation,” *IEEE Int’l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[20] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, “Adaptive step-size parameter control for real world blind source separation,” *IEEE Int’l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[21] F. Jelinek, “Statistical methods for speech recognition,” *MIT Press, Cambridge*, 1997.

[22] R. Gomez, T. Kawahara, K. Nakamura, and K. Nakadai, “Multi-party human-robot interaction with distant-talking speech recognition,” *ACM/IEEE Int’l Conf. Human-Robot Interaction (HRI)*, pp. 439–446, 2012.

[23] R. Gomez, K. Nakamura, and K. Nakadai, “Automatic distance compensation for robust voice-based human-computer interaction,” *Int’l Conf. Machine Intelligence*, 2013.

[24] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from a single depth image,” *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011.

[25] T. Tung and T. Matsuyama, “Human motion tracking using a color-based particle filter driven by optical flow,” *European Conf. Computer Vision Workshop (ECCV)*, 2008.

[26] P. Viola and M. Jones, “Robust real-time object detection,” *Int’l Journ. Computer Vision (IJCV)*, 2001.

[27] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *European Conf. Computer Vision (ECCV)*, pp. 484–498, 1998.

[28] G. Fanelli, T. Weise, J. Gall, and L. V. Gool, “Real time head pose estimation from consumer depth cameras,” *In Proc. DAGM*, 2011.

[29] S. Xu, H. Jiang, and F. C. Lau, “User-oriented document summarization through vision-based eye-tracking,” *13th ACM Int’l conf. Intelligent User Interfaces*, 2008.

[30] L. Feng, Y. Sugano, T. Okabe, and Y. Sato, “Inferring human gaze from appearance via adaptive linear regression,” *IEEE Int’l Conf. Computer Vision (ICCV)*, 2011.

[31] T. Tung, S. Nobuhara, and T. Matsuyama, “Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo,” *IEEE Int’l Conf. Computer Vision (ICCV)*, 2009.

[32] T. Kawahara, S. Hayashi, and K. Takanashi, “Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations,” *Conf. Interspeech*, 2013.

[33] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, “Dynamic textures,” *Int’l Journ. Computer Vision (IJCV)*, vol. 51, no. 2, 2003.

[34] A. Ravichandran, R. Chaudhry, and R. Vidal, “View-invariant dynamic texture recognition using a bag of dynamical systems,” *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009.

[35] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, “Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions,” *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009.

[36] T. Tung and T. Matsuyama, “Intrinsic characterization of dynamic surfaces,” *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2013.

[37] ——, “Timing-based local descriptor for dynamic surfaces,” *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014.

[38] A. B. Chan and N. Vasconcelos, “Mixtures of dynamic textures,” *IEEE Int’l Conf. Computer Vision (ICCV)*, 2005.

[39] J. Koenderink and A. van Doorn, “Surface shape and curvature scales,” *Image and Vision Computing*, 1992.

[40] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, “Performance capture from sparse multi-view video,” *ACM Trans. Graphics*, vol. 27, no. 3, 2008.

[41] F. David, R. Pisani, and R. Purves, “Statistics, fourth edition,” *W. W. Norton & Company*, 2007.

[42] T. Tung and T. Matsuyama, “Topology dictionary for 3d video understanding,” *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, no. 8, pp. 1645–1657, 2012.

Tony Tung received the M.Sc. degree in Physics and Computer Science from the Ecole Nationale Supérieure de Physique, France, in 2000, and the Ph.D. degree in Signal and Image processing from the Ecole Nationale Supérieure des Télécommunications de Paris, France, in 2005. Since 2010, he has been an Assistant Professor at Kyoto University, working jointly at the Graduate School of Informatics and at the Academic Center for Computing and Media Studies. His research interests include computer vision, pattern recognition, shape modeling, and multimodal interaction.

Randy Gomez received the M.Eng.Sci. degree in electrical engineering from the University of New South Wales (UNSW), Sydney, Australia, in 2002 and the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (Shikano Laboratory), Nara, Japan, in 2006. His research interests include robust speech recognition, speech enhancement, acoustic modeling and adaptation, computational scene analysis and human robot interaction. Currently, he is a scientist at Honda Research Institute Japan.

Tatasuya Kawahara (M’91-SM’08) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Affiliated Professor in the Graduate School of Informatics, Kyoto University.

Takashi Matsuyama received B. Eng., M. Eng., and D. Eng. degrees in electrical engineering from Kyoto University, Japan, in 1974, 1976, and 1980, respectively. He is currently a professor in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. His research interests include knowledge-based image understanding, computer vision, 3D video, human-computer interaction, and smart energy management.