Group Dynamics and Multimodal Interaction Modeling using a Smart Digital Signage

Tony Tung, Randy Gomez, Tatsuya Kawahara, and Takashi Matsuyama

Kyoto University,
Academic Center for Computing and Media Studies
and Graduate School of Informatics, Japan
tung@vision.kuee.kyoto-u.ac.jp,
{randy-g,kawahara}@ar.media.kyoto-u.ac.jp, tm@i.kyoto-u.ac.jp

Abstract. This paper presents a new multimodal system for group dynamics and interaction analysis. The framework is composed of a mic array and multiview video cameras placed on a digital signage display which serves as a support for interaction. We show that visual information processing can be used to localize nonverbal communication events and synchronized with audio information. Our contribution is twofold: 1) we present a scalable portable system for multiple people multimodal interaction sensing, and 2) we propose a general framework to model A/V multimodal interaction that employs speaker diarization for audio processing and hybrid dynamical systems (HDS) for video processing. HDS are used to represent communication dynamics between multiple people by capturing the characteristics of temporal structures in head motions. Experimental results show real-world situations of group communication processing for joint attention estimation. We believe the proposed framework is very promising for further research.

1 Introduction

Over the last decades electronic displays have become ubiquitous and have participated in many everyday life activities. Digital advertising displays, video games or poster presentations trigger group discussions which generally contain lots of interactions, and therefore lots of information on human communication and behavior. Here, we present a novel multimodal system to capture and analyze multiple people dynamics and interaction. The system detects and recognizes verbal and non-verbal communication signals, and returns human readable feedbacks on a display screen. Visual information processing is used to detect communication events that are synchronized with audio information (e.g., head motion and speech). The system could potentially be adapted for various applications such as entertainment (multiplayer interactive gaming device), education or edutainment (virtual support for lecturer), medicine, etc.

Multimodal Audio/Video systems designed for human behavior and interaction analysis usually consist of multiple video cameras and microphones placed in a dedicated room, and oriented towards the participants. To date, these systems are still very tedious to setup and often require wearable equipments that

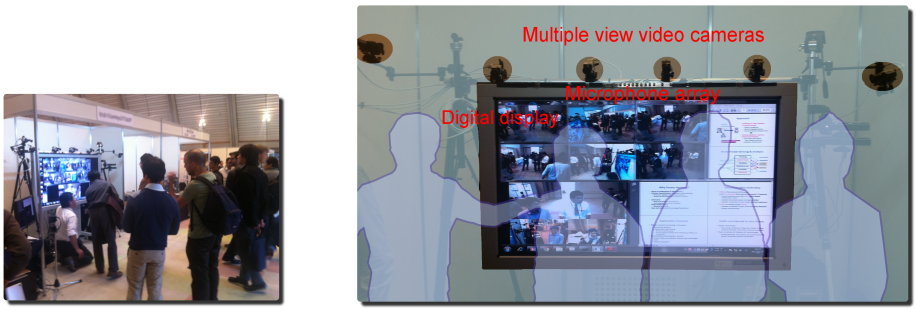

Fig. 1. Smart digital signage tested during poster presentation for group dynamics and multimodal interaction analysis.

prevent them to be used casually or in an uncontrolled environment. Hence, we propose a scalable portable system (i.e., all the devices are transportable while their number can be increased) that employs state-of-the-art techniques in graphics (GPU), vision, and speech processing for multimodal interaction sensing and analysis. Non-verbal signals from head motions are identified and correlated with speech data, and their dynamics are modeled using hybrid dynamical systems (HDS). We show that HDS can be used to obtain temporal structures (i.e., duration, overlaps, etc.) of multimodal events for interaction analysis. The current system has been setup for multiple subjects interacting in front of a large digital display at a short distance, out of the range of consumer depth cameras (see Fig. 1). Real poster presentations as well as casual discussions were captured using the system for joint attention estimation. The next sections present related work, a description of the framework, A/V multimodal interaction modeling using speaker diarization and hybrid linear dynamical systems, experimental results, and a conclusion about our contribution.

2 Related work

Interaction modeling has been a very attractive research topic since decades due to its multidisciplinary aspect. For example, human-to-human and human-computer interaction have been studied in numerous fields of science such as psychology [1], computer graphics [2], communication [3, 4], etc. In group communication, humans use visual and audio cues to convey and exchange information. Hence video and audio data have been naturally extensively used to study human behavior in communication. For example, several corpus such as VACE [5], AMI [6], Mission Survival [7], IMADE [8] were created to capture multimodal signals in multi-party conversation and interaction. Speech is often used to detect dominant speakers based on turn-taking and behavior analysis, while non-verbal cues provide feedbacks to understand communication patterns and behavior subtleties (e.g., smile, head nodding or gaze direction change) and

can be used as back-channels to improve communication [9]. Nevertheless heavy equipments (e.g., headsets) are often required, visual information processing is usually limited (due to video resolution), and no solution is given for automatic multimodal information analysis.

As shown in the literature, the Hidden Markov Models (HMM) are very popular for speech and gesture modeling and recognition [10, 11]. However, limitations lie in the lack of flexibility for timing structure manipulation (e.g., duration of states and transitions), 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. LDS have been applied for dynamic texture modeling [12], facial movement synchronization [13], human action recognition [14], etc. In our framework, we use hybrid dynamical systems (HDS) to model nonverbal behaviors which are synchronized with speech.

To our knowledge no similar framework has been proposed in the literature that aims at multi-people interaction modeling using multimodal (audio and video) signals to study human behavior in group communication (e.g., to detect and analyze joint attention of audience). Other systems using digital signage, like the moodmeter from MIT, usually require only one video camera that performs *only* face detection/classification. Audio is not used and they do not consider human-human interaction. Commercial systems, like Samsung Smart TVs, use single-human gestures as remote control and do not handle interaction between multiple people.

3 Multimodal sensing framework

3.1 Audio/Video system setting

Audio. We employ a hands-free speech communication setup in the capture environment to give subjects more degrees of freedom in interacting with each other. This setup precludes holding or wearing a physical microphone. Although signal-to-noise (SNR) ratio is significantly lower in the hands-free setup as compared to the close-distant talking microphones, we mitigate this issue by using a microphone array. The increase in microphone count results in an improvement of the SNR. In our setup, we use 19-channel microphone array in a linear configuration attached on top of a 65-inch digital display (see Fig 1). Each signal from the microphone is sampled with 16KHz sampling rate, which is sufficient to cover the frequency band of the speech signal.

Video. Multiple video cameras are employed to capture nonverbal communication and interaction between multiple people. 6 HD video cameras are placed on a pole mounted on the display to obtain wide field of view (270 deg) and dense 3D face reconstruction. To keep the design simple, only one PC with a single GPU is used for video capture and processing. Videos are recorded simultaneously in SXGA at 15fps using Point Grey 1394b cameras with wide angle

3.5mm lenses. Note that to date, hardware synchronization of HR cameras with standard depth cameras is still not possible.

3.2 Multimodal signal capture

Audio. Aside from mere convenience, hands-free speech communication through microphone array offers meaningful signal processing tools. Data from the different channels can be processed to suppress contaminants emanating from noisy sources in real environment condition through beamforming [15]. Moreover, microphone array processing can also be used to effectively focus the microphone sensitivity to the party of interest, and further enhance the speech signal. This minimizes cross talk from the other speech sources or unwanted noise coming from the environment. Then, nonlinear processing technique is introduced in which the speech from other sources (other than that of the party of interest) is transformed to noise [16]. For example, in a poster presentation scenario, the party of interest is either the presenter or the audience, thus we transform either one of these to noise and enhance the other. As a result, the processed audio stream contains both the enhanced speech of the party of interest and noise (transformed speech).

Video. The proposed system detects and tracks multiple people faces from multiple views. As we use HD cameras, appearance-based methods return reliable detection results [17]. Face detection is combined with face feature detection (e.g., nose) for the sake of robustness, and computed on GPU to speed up calculations. To achieve simultaneous detections from multiple views with a single GPU, we first build a composite image by concatenation of regions of interests from multiview frames, and then transfer the image to the GPU; e.g., a consumer graphics card (GeForce NVIDIA GTX) can easily handles 3 frames simultaneously in real-time. Our face tracker employs a Bayesian model and online learning for continuous tracking [18]. Here, face feature coordinates, face templates, detection scores, and depth distributions are used as priors to estimate posterior probabilities of face positions. Dense 3D face reconstruction from stereo and point cloud noise removal using spatio-temporal joint bilateral filtering are also computed online (see Fig.2). Head pose can therefore be estimated by a geometrical approach (model fitting) to derive head motion and gaze direction. See Fig. 3 (bottom) for an overview of the process.

4 Multimodal interaction dynamics

Temporal structures in speech and head motion play a crucial role in natural human communication. While speech processing from audio data allows speaker turn diarization, dynamic features from visual information processing can be modeled using an interval-based representation of hybrid dynamical systems (IHDS) that model human communication event dynamics [13]. The proposed strategy allows the identification of behavior patterns in multimodal interaction such as when joint attention occurs (see Fig. 3).

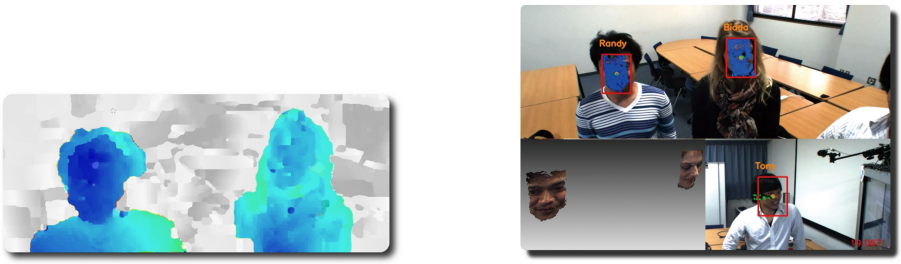

Fig. 2. Video processing: (Left) Real-time depth map from stereo; (Center and Right) multi-people face detection, tracking, and 3D face for head pose estimation.

4.1 Speaker diarization

Diarization of speaker turns involves classifying one speaker from the other: e.g. in the case of a poster presentation, identifying the presenter-audience turn. When considering speech as the mode of input in the diarization task, the performance of the system primarily depends on separating the presenter's speech from that of the audience. However, separation is not straightforward since speech itself shares a common subspace even when spoken by different people. This is the reason why speech recognition technology is usually speaker-independent (e.g. speech from different people can still be recognized even if not enrolled during training). Thus, the technique in the microphone array processing circumvents this problem by treating the speech-speech classification approach into speech-noise classification.

We note that speech and noise subspaces are distinct, which minimizes classication ambiguity. In our framework, we design two Gaussian mixture model (GMM) classiers (e.g., λ_S for speech and λ_N for noise). Depending on the size of the training data, Gaussian components are increased to improve subspace discrimination. This process is terminated when the classification accuracy reaches the saturation value. Specically, we use 256 Gaussian components for each model. The two GMMs are trained by means of Expectation-Maximization [19]. The microphone array-processed data is windowed using a 25-ms frame. Then, mel cepstrum, energy and delta energy features are extracted, which are used in the training phase. These features suciently capture the relevant speech information with reduced dimensionality. In the actual diarization scheme shown inFig. 3 (top), identication of the speaker turn is implemented by processing the 19-channel mic array signals resulting to $\bar{x}$. The processed data contain the enhanced speech (party of interest) and noise (unwanted party). Then, likelihood score is evaluated using the mic array-processed stream against the 2 GMMs (λ_S and λ_N). Finally, the GMM that results to a higher likelihood score is selected as the corresponding class.

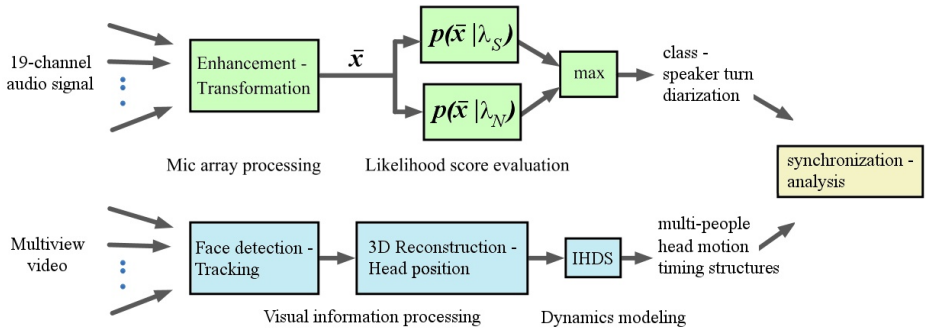

Fig. 3. Processing scheme for multimodal interaction analysis.

4.2 Hybrid linear dynamical system

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.

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

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

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 null averages and Q_i and R as covariances respectively. In order to control the system state changes between two events, an FSM having a discrete set of states $\mathcal{S} = \{s_i\}_{i=1\dots N}$ 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 [13].

Interval representation. Interval-based representation of HDS (IHDS) is used to describe event timing structures (see Fig. 3 (bottom)) and can be used for event classification or recognition [13]. 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 behavior can then be modeled using an IHDS, similar to a musical score where $\{I_k\}$ are notes and N is the scale. 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 $\{y(t)\}$.

Interaction analysis. 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:

$$Z(Y_k, Y_{k'}) = Pr(\{b_k - b'_k = \Delta_b, e_k - e_{k'} = \Delta_e\}_{\mathcal{I}} | \{[b_k, e_k] \cap [b_{k'}, e_{k'}] \neq \emptyset\}_{\mathcal{I}}]), \quad (2)$$

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 (especially where $|b_k - b_{k'}| < 1s$). If $\{(b_k - b_{k'}) \to 0\}$ and $\{(e_k - e_{k'}) \to 0\}$, then all pairs of overlapping intervals are synchronized.

5 Experimental results

To assess the performance of our framework, the setup was tested in real-world situations such as a conference hall and a meeting room. Despite fairly cluttered backgrounds and various illumination conditions, the system was effective and poster presentations as well as casual discussions between 3-4 people were held to evaluate joint attention of subjects from multimodal event interaction analysis (see Fig. 1,2). Audio and multiview video are captured simultaneously, and an offline process outputs multimodal interaction levels within seconds.

In Figure 4, we show the results obtained with two sequences with some ground-truth hand-made annotations : a) a poster presentation involving a presenter and 2-people audience (2000 frames), and b) a casual discussion between 3 subjects commenting photos displayed on the screen (2500 frames). Head motion dynamics were modeled using HDS from head positions (x, y) (see plots). We show interval-based representations of HDS model states (IHDS) with $N = 4$ modes. Here, state changes correspond to head motions (e.g., turning, nodding, etc.). Presenter in a) and Subject 1 in b) who were closer to the display did numerous head movements towards the screen and other subjects (26.5/min v 32.3/min). In a) Audience 2 produced much more nonverbal communication signals than Audience 1 (32/min v 19.5/min), whereas in b) Subject 2 and Subject

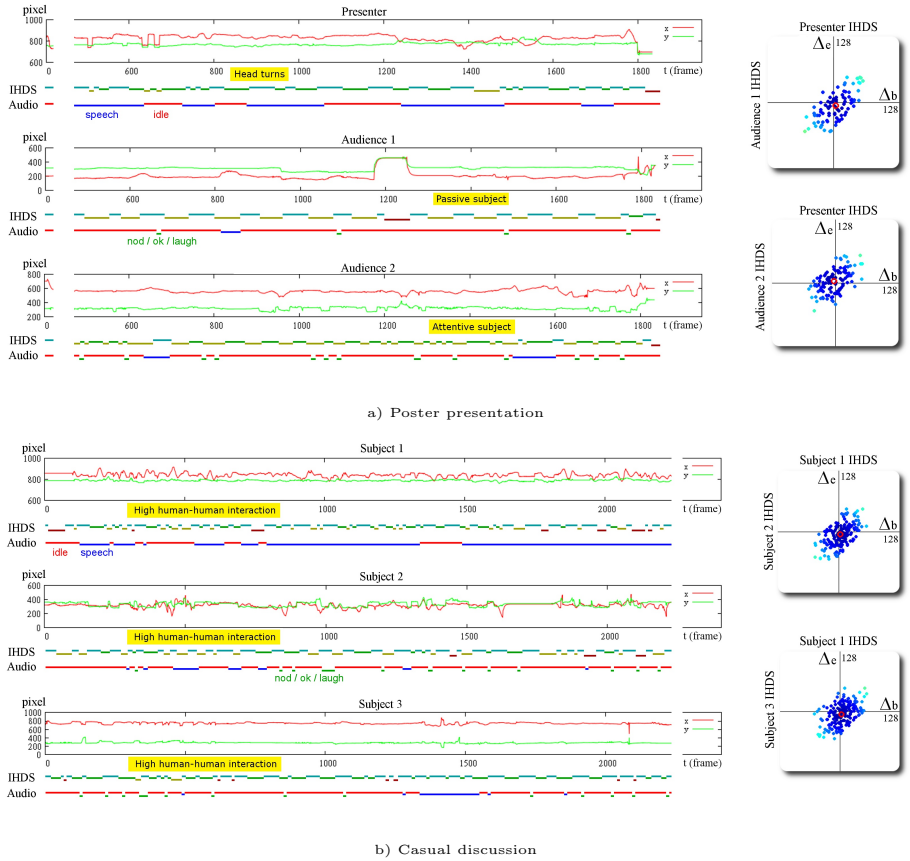

Fig. 4. Group dynamics and multimodal interaction modeling for: a) Poster presentation and b) Casual discussion. From the top: head position (x, y) in pixels, IHDS modeling with 4 modes, and speaker diarization (red: idle, blue: speech, green: nod/ok/laugh). Right: IHDS synchronization distributions.

3 performed similarly (25.3/min v 27.3/min). As can be observed, interactions between participants were more frequent during the casual discussion in b). Also, face tracking of Audience 1 in a) was lost around frame 1200 during the processing due an implementation issue. Nevertheless the unexpected tracking behavior has been successfully identified as a separate state by the HDS model.

Signal synchronizations (see right in a) and b)) show all synchronized interval disparities between Presenter and Audience 1 and 2, and between Subject 1 and Subject 2 and 3 respectively. The temporal difference distributions have a maximum $|\Delta b|$ and $|\Delta e|$ of 60f (4s). The centers of the distributions are close to the center (red circle), meaning mere synchronization. Note that in the context of poster presentation, the position of the subjects does not change a lot. Hence,

we could consider global head motions without cancelling the body motions. Therefore, reactions to signals from the head include as well reactions from body motions (e.g., body translation can create reaction).

In both scenarios, one participant is more active than the others: the Presenter in a) with 62.5% of speech, and the Subject 1 in b) with 73.5%. (Audience 1 and 2 have 3.1% and 9.4% of speech respectively, and Subjects 2 and 3 have 14.2% and 12.3.) Hence, we propose to use Eq. 4.2 to evaluate joint attention of the other participants by analyzing multimodal interactions and measuring interaction levels. Figure 5 shows interaction levels between the main speaker and each participants, i.e., the number of reactions with respect to reaction time. In a) and b), we show: (Left) head reactions in response to audio stimuli for all participants (main speaker included), and (Right) head reactions of participants other that the main speaker in response to visual stimuli from him. In a), we can see again that Audience 2 has much more reactions than Audience 1 for both audio and visual stimuli. More reactions are found with the visual stimuli: Audience 2 (46 reactions per minute) v Audience 1 (33rpm), The audio stimuli return: Audience 2 (8rpm) v Audience 1 (5rpm). In b), the number of reactions are similar, showing equal interaction level between Subject 2 and 3. Audio: 13rpm v 11rpm, and video: 83rpm v 87rpm. 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 we can observe that reaction times of Audience 2 and Subject 2 are very good, which is reconfirmed by checking the videos.

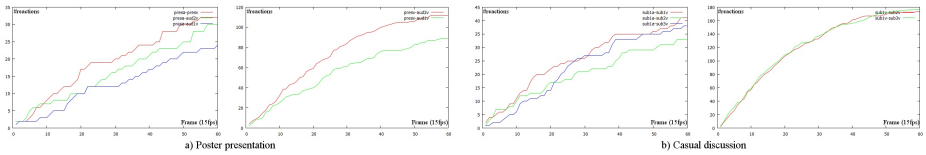

Fig. 5. Multimodal interaction level with respect to reaction time: a) Poster presentation, and b) Casual discussion.

6 Conclusion

This paper presents a new framework for group dynamics and multimodal interaction modeling. The proposed system is portable and scalable and consists of a smart digital signage display equipped with a mic array and multiview video cameras. We capture multiple human interaction events and analyze them automatically using audio and visual information processing. We show that communication dynamics can be used to estimate joint attention using an interval-based representation of hybrid dynamical systems and speaker turn diarization. To our knowledge, no similar framework has been proposed yet.

Acknowledgments. This work was supported in part by the JST-CREST project "Creation of Human-Harmonized Information Technology for Convivial Society", and the Japan Society for the Promotion of Science (Wakate-B No. 23700170). The authors would like to thank Dr. Hiroaki Kawashima for his inspirational work on IHDS.

References

1. Newcomb, T.M., Turner, R.H., Converse, P.E.: Social psychology: The study of human interaction. Routledge and Kegan Paul (1966)
2. Cassell, J., Vilhjálmsson, H., Bickmore, T.: Beat: the behavior expression animation toolkit. SIGGRAPH (2001)
3. Buchanan, M.: Secret signals. Nature (2009)
4. Pentland, A.: To signal is human. American Scientist (2010)
5. Chen, L., Rose, R., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R., Huang, T.: Vace multimodal meeting corpus. (2006)
6. Poel, M., Poppe, R., Nijholt, A.: Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction. FG (2008)
7. Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of concensus decision making meetings. Language Resources and Evaluation (2007) 409–429
8. Sumi, Y., Yano, M., Nishida, T.: Analysis environment of conversational structure with nonverbal multimodal data. ICMI-MLMI (2010)
9. White, S.: Backchannels across cultures: A study of americans and japanese. Language in Society **18** (1989) 59–76
10. Rabiner, L.R.: A tutorial on hidden markow models and selected applications in speech recognition. IEEE **77** (1989) 257–286
11. Liu, C.D., Chung, Y.N., Chung, P.C.: An interaction-embedded hmm framework for human behavior understanding: With nursing environments as examples. IEEE Trans. Information Technology in Biomedecine **14** (2010) 1236 – 1246
12. Doretto, G., Chiuso, A., Wu, Y., Soatto, S.: Dynamic textures. IJCV **51** (2003)
13. Kawashima, H., Matsuyama, T.: Interval-based modeling of human communication dynamics via hybrid dynamical systems. NIPS Workshop on Modeling Human Communication Dynamics (2010)
14. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. CVPR (2009)
15. Jani, E., Heracleus, P., Ishi, C., Nagita, N.: Joint use of microphone array and laser range finders for speaker identification in meeting. Japanese Society for Artificial Intelligence (2011)
16. Gomez, R., Lee, A., Saruwatari, H., Shikano, K.: Robust speech recognition with spectral subtraction in low snr. Int'l Conf. Spoken Language Processing (2004)
17. Viola, P., Jones, M.: Robust real-time object detection. IJCV (2001)
18. Perez, P., C, C.H., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. ECCV (2002)
19. Gomez, R., Kawahara, T.: Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. IEEE Trans. Audio, Speech and Language Processing (2010)