Multi-party Human-Machine Interaction Using a Smart Multimodal 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@ar.media.kyoto-u.ac.jp,
kawahara@ar.media.kyoto-u.ac.jp, tm@i.kyoto-u.ac.jp

Abstract. In this paper, we present a novel multimodal system designed for smooth multi-party human-machine interaction. HCI for multiple users is challenging because simultaneous actions and reactions have to be consistent. Here, the proposed system consists of a digital signage or large display equipped with multiple sensing devices: a 19-channel microphone array, 6 HD video cameras (3 are placed on top and 3 on the bottom of the display), and two depth sensors. The display can show various contents, similar to a poster presentation, or multiple windows (e.g., web browsers, photos, etc.). On the other hand, multiple users positioned in front of the panel can freely interact using voice or gesture while looking at the displayed contents, without wearing any particular device (such as motion capture sensors or head mounted devices). Acoustic and visual information processing are performed jointly using state-of-the-art techniques to obtain individual speech and gaze direction. Hence displayed contents can be adapted to users' interests.

Keywords: multi-party, human-machine interaction, digital signage, multimodal system

1 Introduction

Over the last decade smart systems for multi-party human-machine interaction have become ubiquitous in many everyday life activities (e.g., digital advertising displays, video games or poster presentations). Here, we present a novel multimodal system that is designed for smooth multi-party human-machine interaction. The system detects and recognizes verbal and non-verbal communication signals from multiple users, and returns feedbacks via a display screen. In particular, visual information processing is used to detect communication events that are synchronized with acoustic information (e.g., head turning and speech). We believe the system can potentially be adapted to 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 participants. To date, these systems are still very tedious to setup and often require wearable equipments that prevent them to be used casually or in an uncontrolled environment. Here, the proposed system is portable and scalable. It consists of a digital signage or large display equipped with multiple sensing devices spaced on a portable structure: a microphone array, 6 HD video cameras, and two depth sensors. The display is used to show various contents, such as poster presentations, web browsers, photos, etc. On the other hand, 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). We tested the system with real poster presentations as well as casual discussions. The next sections present related work, description of the framework, A/V multimodal data processing, application to multi-party interaction, and conclusion.

2 Related Work

Human-to-human and human-computer interaction have been studied in numerous fields of science such as psychology, computer graphics, communication, 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 [Chen et al., 2006], AMI [Poel et al., 2008], Mission Survival [Pianesi et al., 2007], IMADE [Sumi et al., 2010] 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 [White et al., 1989]. Nevertheless heavy equipments (e.g., headsets) are often required, and visual information processing is usually limited (due to video resolution). Other systems using digital signage (like the moodmeter from MIT) usually use only one video camera that performs only face detection/classification. Acoustic is not used and they do not consider human-machine interaction. Commercial systems, like Samsung Smart TVs, use single-human gestures as remote control and do not handle interaction with multiple people. To our knowledge, no framework has been proposed in the literature that aims at multi-people interacting with a digital signage using multimodal (audio and video) signals. Note that in [Tung et al., 2012], the authors introduce a multimodal interaction model for multi-people interaction analysis that can be used with the system presented in this paper. As well, let us cite [Tung and Matsuyama, 2012] in which the authors present 3D scene understanding using data from multiview video capture.

3 Smart multimodal system configuration

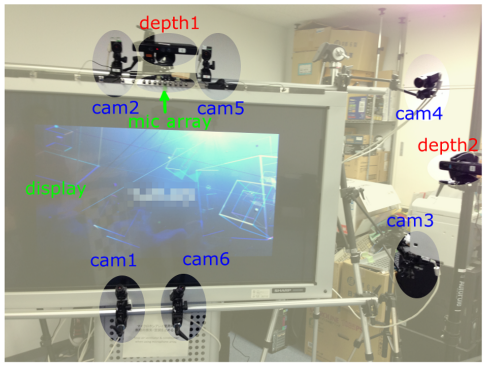
Audio. Microphone array is employed as audio capturing device. The spatial arrangement of the microphone sensors enable the system to localize different sound sources: the speech signal that originates from the individual participants and the signals that come from other sources (i.e., background noise). The microphone array system provides a hands-free audio capture mechanism in which the participants are not constrained to using wired microphones that limits movements, paving the way towards free flowing interaction. The power of the captured audio signals from the distant sources are improved by increasing the number of microphone sensors in the array, in this case, a reliable level of speech power is achieved which is very important in the speech recognition system. The microphone array is configured linearly with 19 microphone sensors and placed on top of a 65-inch display display (see Fig. 1) and the sampling rate is 16KHz. The smart audio processing in a poster presentation scenario (i.e., participants: presenter and audience) is described in Sec. 4.

Video. Multiple video cameras are employed to capture nonverbal communication and interaction between multiple people. In order to capture multiple views of subjects standing in front of the system, 6 HD video vision cameras (from Point Grey) are spaced on a portable structure made of poles and mounted around a display. In the context of poster presentation, sensing devices are placed at on one side of the display to capture a presenter, and at the center to capture the audience (e.g., two or three people). We place 3 cameras with 3.5mm lenses on top of the display (two at the center, one on the side) to obtain wide field-of-view (150 deg), perform video capture in UXGA at 30 fps, and 3D reconstruction from stereo. As well, 3 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 30 fps. As well, two depth sensors (MS Kinect) are placed on top of the screen (one at the center, one on the side) and capture depth map videos in VGA at 30 fps. To make the system easily transportable, only one PC with a single GPU is used for video capture and processing. Fig. 1 shows the current system and multiview video samples.

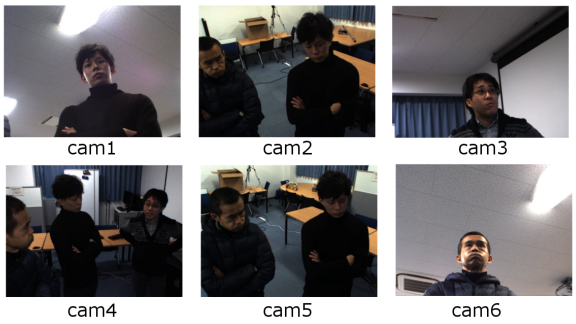
4 Multimodal data processing

Audio. Acoustic signal processing in a multi-party system involves the processing of the data captured from the microphone array and the design of the automatic speech recognition (ASR) system. The multimodal signage application can be used in various scenarios, specifically in this paper we focus on poster presentation that involves the interaction between the presenter and the audience. Audio processing is described as follows:

- Microphone Array Processing
  Assuming that there are 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



a) Mutimodal system



b) Multiview video

Fig. 1. a) Multimodal system setup for smart poster presentation. b) Multiview video capture of one presenter and two attendees.

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 is denoted as $\boldsymbol{o}(\omega) = [o_1(\omega), \cdots, o_M(\omega)]^T$. Microphone array signal processing is described 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 behaviour of the sound signal as it is reflected inside the room enclosure. The background noise is denoted by $\boldsymbol{n}(\omega)$. We note that the both $\boldsymbol{n}(\omega)$ and $\boldsymbol{s}(\omega)$ are assumed to be statistically independent which is usually true in real environment conditions 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 beam

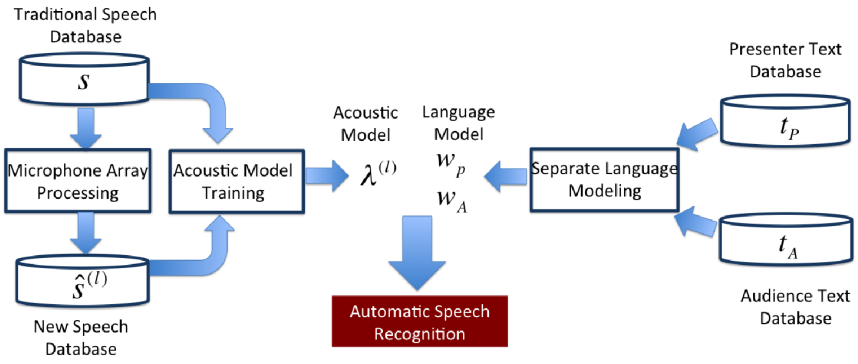


Fig. 2. Robust modeling for automatic speech recognition system.

forming and blind separation [Sawada et al., 2002][Nakajima et al., 2008]. The separated signal is denoted as $\hat{s}^{(l)}(\omega)$.

– Context and Models for ASR

We trained separate acoustic models based on different user profiles. During the early stage of the design process, roles of the participants are profiled in advance. For example, if the the system is used for poster presentation, participants are classified as presenter, audience, etc. The task of the participants are known right from the very beginning; the presenter usually has a script which is used to discuss the content of the presentation displayed on the digital signage. In most cases, the signage itself contains information of the presenters talk (i.e., text). Moreover, the audience is presumed to ask questions, and because of this prior information, it is safe to assume that the conversation dynamics between the participants are readily available. Consequently, depending on the directivity of the speaker we can re-train the acoustic model to $\lambda^{(l)}$ with data processed using the sound separation mechanism discussed above. This will improve performance as opposed to solely using the baseline model λ trained from the traditional speech database because the actual location (l) of the speakers are considered in the former. Separate language models can also be modeled for the presenter and the audience, respectively. By using the corresponding text data that are unique to the conversation dynamics to each of the participant class, separate language models are trained: w_p (for presenter) and w_a (for audience). Training procedure that involves context is shown in Fig. 2. In our method, we are able to break down the generic approach in the automatic speech recognition system (ASR). By incorporating context with respect to the microphone array processing reflective of the participant's position and the separate language models, the system can automatically switch parameters that are appropriate to the current scenario. In the end, robustness is achieved.

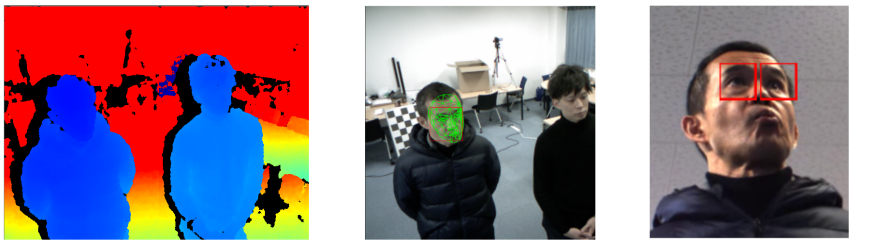


Fig. 3. a) Depth map from frontal depth sensor. b) Face feature tracking using depth and color information (by Kinect SDK). c) Eye localization after registration to HD video camera 6.

Video. Visual information processing is achieved using the combination of depth sensors that deliver depth maps of the scene in VGA resolution, and multiple video cameras that capture the scene in higher resolution (UXGA and SVGA). All multiview video cameras are geometrically calibrated using standard methods (i.e., with a chessboard) and synchronized by software trigger. The depth sensors (IR cameras) are also calibrated with the HD video cameras. Depth data serve for human body-part identification using approaches based on discriminative random regression forests that efficiently label different body regions [Shotton et al.,2011]. As well, head pose and face feature tracking can be performed from depth and color information of the depth sensors using state-of-the-art techniques [Cootes et al., 2006][Viola et al., 2001][Fanelli et al., 2011]. As the resolution of color cameras integrated in currentconsumer depth sensors are usually too poor to provide accurate gaze estimation, HD video cameras placed below the screen are used instead for accurate gaze direction estimation [Xu et al., 2008][Feng et al., 2011]. In practice, as shown in Fig. 3 we extract regions of interest (e.g., faces and eyes), and register HD video frames to the depth maps. Compared to prior multimodal setup [Tung et al., 2012], the proposed design is able to provide more accurate head pose and gaze estimation based on eye detection. Note that, as in [Tung et al., 2012] HD video cameras placed on top of the screen provide rough depth maps in real-time (using multiview stereo reconstruction), which can be merged with data from the depth sensors for further depth accuracy enhancement. Furthermore, head and mouth positions are used in the speech processing described above for better diarization.

5 Applications: Multi-party Multimodal Interaction

In this work, we combine acoustic and video information for seamless interaction with smart display. Suppose that a display contains text data localized into several regions according to visual coordinate locations. The generic word vocabulary which is composed of the total text is broken down into sub-regions,

Fig. 4. When multiple browsers are opened simultaneously, it is not trivial for hands-free voice systems to understand what applications users are focusing on. The proposed system uses gaze and speech processing to determine what actions to perform.

corresponding to the several visual coordinate locations. Using the localized vocabulary (within a particular region), a new set of language model is updated for each region. This allows us to dynamically select active regions on the display based on the gazes of the participants. The system dynamically switches language models for speech recognition, reflective of the active region. This strategy minimizes the confusion in speech recognition when displaying multiple contents, as a particular word entry may occur several times in different locations within the whole display. For example, as illustrated in Fig. 4 when multiple browsers are opened simultaneously, it is not trivial for hands-free voice systems to understand what applications users are thinking about when simple actions are requested (e.g., close browser, search in window, etc.).

Switching to active regions based on the users' gaze narrows down the size of the vocabulary as defined by each active region. In effect, the system benefits from both acoustic and visual information in improving overall system performance. Furthermore, the system can be used as a smart poster for automatic presentation to a group of people. By capturing and analyzing gaze directions and durations, attention or interest levels can be derived for each portion of the presentation (see [Tung et al., 2012]). Hence, when joint-attention is detected (e.g., when two subjects are looking at the same region), the system can automatically highlight the region of interest. We believe the proposed system can be used for numerous applications, such as education, medicine, entertainment, and so on.

6 Conclusion

We present a novel multimodal system that is designed for smooth multi-party human-machine interaction. The system detects and recognizes verbal and non-verbal communication signals from multiple users, and returns feedbacks via a

display screen. In particular, visual information processing is used to detect communication events that are synchronized with acoustic information (e.g., head turning and speech). To our knowledge, no similar setup has been proposed yet in the literature.

Acknowledgment. This work was supported in part by the JST-CREST project Creation of Human-Harmonized Information Technology for Convivial Society. The authors would like to thank Hiromasa Yoshimoto for his work on system development and data capture.

References

[Chen et al., 2006] 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. MLMI, LNCS 3869, Springer(2006)

[Cootes et al., 2006] Cootes, T.F., Edwards, G. J., Taylor, C. J.: Active appearance models. ECCV, 2:484498, 1998

[Fanelli et al., 2011] Fanelli, G., Weise, T. , Gall, J., Van Gool, L.: Real Time Head Pose Estimation from Consumer Depth Cameras. DAGM (2011)

[Feng et al., 2011] Feng, L., Sugano, Y., Okabe, T., Sato, Y.: Inferring human gaze from appearance via adaptive linear regression. ICCV (2011)

[Gomez and Kawahara, 2010] Gomez, R., Kawahara, T.: Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. IEEE Trans. Audio, Speech and Language Processing, 18(7), 1708-1716 (2010)

[Nakajima et al., 2008] Nakajima, H., Nakadai, K., Hasegawa, Y., Tsujino, H.: Adaptive Step-size Parameter Control for real World Blind Source Separation. ICASSP (2008)

[Pianesi et al., 2007] Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of concensus decision making meetings. Language Resources and Evaluation, 409429 (2007)

[Poel et al., 2008] Poel, M., Poppe, R., Nijholt, A.: Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction. FG (2008)

[Sawada et al., 2002] Sawada, H., Mukai, R., Araki, S., Makino, S.: Polar coordinate based nonlinear function for frequency-domain blind source separation. ICASSP (2002)

[Shotton et al.,2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T. , Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from a Single Depth Image. CVPR (2011)

[Sumi et al., 2010] Sumi, Y., Yano, M., Nishida, T.: Analysis environment of conversational structure with nonverbal multimodal data. ICMI-MLMI (2010)

[Tung and Matsuyama, 2012] Tung, T., Matsuyama, T.: Topology Dictionary for 3D Video Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 34(8), 1645-1657 (2012)

[Tung et al., 2012] Tung, T., Gomez, R., Kawahara, T., Matsuyama, T.: Group Dynamics and Multimodal Interaction Modeling using a Smart Digital Signage. ECCV, Ws/Demos, LNCS 7583, Springer (2012)

[Viola et al., 2001] Viola, P., Jones, M.: Robust real-time object detection. IJCV (2001)

[White et al., 1989] White, S.: Backchannels across cultures: A study of americans and japanese. Language in Society 18, 5976 (1989)
[Xu et al., 2008] Xu, S., Jiang, H., Lau, F. C.: User-oriented document summarization through vision-based eye-tracking. 13th ACM Int'l conf. Intelligent User Interfaces (2008)