

COLLABORATIVE CAPTURING AND DETECTION OF HUMAN INTERACTIONS IN MEETINGS

Zhiwen Yu, Hideki Aoyama, Motoyuki Ozeki, Yuichi Nakamura¹

Abstract

This paper proposes a collaborative approach to capture and detect human interactions in meetings by employing multiple sensors, such as video cameras, microphones, and motion sensors. Different from physical interactions (e.g. turn-taking and addressing), the human interactions here are incorporated with semantics, i.e. user intention or attitude towards a topic. The contexts used for interaction detection include head gesture, attention from others, speech tone, speaking time, interaction occasion, and information about previous interaction. The support vector machines (SVM) classifier is adopted to recognize human interactions based on these features. The experimental results verified the proposed approach.

1. Introduction

Meetings are important events in our daily life. People are usually not able to attend all relevant meetings or to remember all the important information produced in meetings. This precipitates the advent of smart meeting that automatically records a meeting and analyzes the generated audio-visual content for future viewing [9].

While most of current smart meeting systems analyze the meeting content for understanding *what* conclusion was made, it is more interesting and important to know *how* a conclusion was made, for example, did all members agree on the outcome, who did not give his opinion, who spoke a little or a lot, etc. Such kind of group social dynamics can be useful for determining whether the meeting was well organized and whether the conclusion was rational. Human interaction plays an important role in understanding this communicative information. Different from physical interactions (e.g. turn-taking and addressing), the human interactions here are defined as behaviors among meeting participants with respect to the current topic, such as proposing an idea, giving some comments, expressing positive opinion, and requesting information. When incorporated with semantics (i.e. user intention or attitude towards a topic), interactions are more meaningful in understanding conclusion drawing and meeting organization. The interactions can be further used for mirroring group activity for the purpose of group monitoring of the way that it operates.

Numerous research has been done in detecting interactions in meetings. AMI project [3] deals with interaction issues including turn-taking, gaze behavior, influence and talkativeness. Stiefelwagen et al [6] propose an approach for the estimation of who was talking to whom based on tracked head poses of the participants. Sumi et al [7] analyze user interactions (e.g. gazing at an object, joint attention, and conversation) during poster presentation in an exhibition room. The above systems mainly focus on detecting physical interactions between participants without any relations with topics. Hence they can not clearly determine participant's attitude or role in a topic discussion.

In this paper, we propose a collaborative approach to capture and detect human *semantic* interactions in meetings by utilizing a variety of context, such as head gesture, attention from others, speech tone, speaking time, occasion of interaction, and information about previous interaction.

¹ The authors are at the Kyoto University, Japan. Corresponding author: Zhiwen Yu, yu@ccm.media.kyoto-u.ac.jp

The context information is gathered through multiple sensors (e.g. video cameras, microphones, and motion sensors). We adopt support vector machines (SVM) [8] classifier to recognize human interaction based on these features.

The rest of this paper is organized as follows. Section 2 presents the capturing environment, device and method. The human interaction detection is described in Section 3. Section 4 presents our preliminary evaluation results. Finally, we conclude the paper and discuss future work in Section 5.

2. Interaction Capturing

Figure 1 shows the setting of our capturing system. The overview of the capturing environment is shown as Figure 1a while Figure 1b shows a participant who wears motion sensor and microphone.

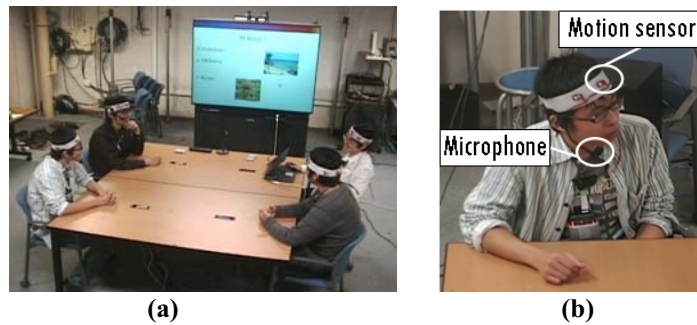


Figure 1. Capturing system: (a) the overview of the capturing environment; (b) a participant who wears motion sensor and microphone

Six video cameras are deployed in our capturing system. Four of them are used for capturing upper body of four participants (breast shot), one camera is employed for recording presentation slides (screen shot), and the other one is utilized for capturing the overview of the meeting including all participants (overview shot). The video signal from each of the camera is input into an encoder board and stored in the format of MPEG2/PS, with frame size as 720*480, frame rate as 29.97 fps, and bit rate as 2Mbps.

To capture the speech pattern during interaction, a head-mounted microphone is attached to each participant. The meeting global sound was mixed with the four individual audio data. The audio signal (MPEG1-Layer II 48000Hz) is also imported into the encoder board in sync with the video signal. The computers attached with the encoder boards are synchronized with each other by NTP.

We use an optical motion capture system (PhaseSpace IMPULSE, <http://www.phasespace.com/>) for head tracking. It mainly consists of three parts: LED module, camera, and server. The tracking system uses multiple CCD cameras for three dimensional measuring on LED tags (sensors) and obtains their exact position. We put six LED tags around the head of each person. The LED tags are scanned 30 times per second, and the position data can be obtained in real-time with less than 10ms latency. Through the three dimensional position data, head gesture (e.g. nodding) and face orientation can be detected.

3. Interaction Detection

3.1. Interaction Type

The various interactions imply different user roles, attitudes, and intentions about a topic during a discussion. The interaction type definition naturally varies according to its usage. We create a set of human interactions that includes seven categories: *propose*, *comment*, *acknowledgement*,

requestInfo, *askOpinion*, *posOpinion*, and *negOpinion*. The detailed meanings are described as: *propose* – a user proposes an idea with respect to a topic; *comment* – a user gives comments on a proposal; *acknowledgement* – a user confirms someone else’s comment or explanation, e.g. “yeah”, “uh huh” and “OK”; *requestInfo* – a user requests information about a proposal; *askOpinion* – a user asks someone else’s opinion about a proposal; *posOpinion* – a user expresses positive opinion, i.e. follow a proposal; and *negOpinion* – a user expresses negative opinion, i.e. against a proposal. Although the Speech Act Theory [5] determines similar semantics based on utterances, we adopt a multimodal approach by utilizing a variety of contexts (e.g. head gesture and face orientation).

3.2. Context Extraction

The context used in our interaction detection includes head gesture, attention from others, speech tone, speaking time, interaction occasion, and information about previous interaction.

Head gesture (e.g. nodding) is very common and used often in detection of human response (acknowledgement or agreement). We determine nodding through the vertical component of the face vector calculated from the position data.

Attention from others is an important determinant of human interaction. For example, when a user is proposing some idea, he is usually being looked at by most of the participants. Attention from others can be treated as how many persons looking at the target user during the interaction. Thus the problem can be roughly turned into detection of face orientation. We measure the angle between the reference vectors (that are from the target person’s head to the other persons’ head) and the target user’s real face vector (calculated from the position data). The face orientation is determined as the one whose vector makes the smallest angle.

Speech tone refers to whether a statement is a question or a normal one. Speaking time is another important indicator in detection the type of human interaction. When a user puts forward a proposal, it usually takes relatively long time. But it takes short time when he gives an acknowledgement or asks a question. The contexts of speech tone and speaking time are automatically detected by using Julius speech recognition engine (<http://julius.sourceforge.jp/>). Julius segments input sound data into speech durations by silence interval longer than 0.3s. We classify segments into question or non-question sentences by using speech’s pitch pattern based on Hidden Markov Models [4] trained with each person’s speech data. The speaking time is derived from the duration of a segment.

The interaction occasion has two values: spontaneous and reactive. The former means the interaction is initiated by the person spontaneously (e.g. proposing an idea or asking a question). The latter denotes the interaction is triggered as response to another interaction. Limited by resource and time, we manually label this feature in our current system.

The type of previous interaction also plays an important role in detecting the current interaction. It is intuitive that there are certain patterns or flows frequently appear in meeting discussion. For instance *propose* and *requestInfo* are usually followed by the interaction of *comment*. This context can be obtained from the recognition result of its previous interaction.

3.3. Interaction Recognition Based on Support Vector Machines

We adopt support vector machines (SVM) classifier for interaction recognition. SVM has been proven to be powerful in classification problems and often achieve higher accuracy than other pattern recognition techniques. It is famous for its strong ability in separating hyper-planes of two or high dimensions. Please refer to [2] for more details about SVM classifier and its inputs/outputs. The LIBSVM library [1] is utilized in our system for classifier implementation. The meeting

content is first segmented into a sequence of interactions. Some samples are selected and fed to SVM as training data, while others are used as testing set.

4. Evaluation Results

The recognition rate is adopted to evaluate the accuracy of our detection mechanisms. It is the ratio between the number of correctly recognized objects and the total number of objects. We first evaluate the performance of context extraction, i.e. the detection accuracy of head gesture (nodding), attention from others (face orientation), and speech tone (question or normal). The results are reported in Table 1. We can see that the accuracy of nodding detection is the highest (76.4%). The recognition rate of speech tone is a little low, but still acceptable.

We further evaluate the interaction detection by measuring the recognition rate. Through the video and audio data, we manually labeled 406 interactions. 311 of them were chosen as training set and the other 95 interactions were used for testing. Different context sets were configured and given to SVM classifier to test the effect of different features on interaction recognition. The context set configuration and recognition results are presented in Table 2. There are totally five different context sets in this experiment. Set 1 is a complete configuration with all the six categories of context. Set 2, 3, 4, 5 include all contexts except head movement (i.e. head gesture and attention from others), speech feature (i.e. speech tone and speaking time), interaction occasion, or feature of previous interaction respectively. It is shown that with all contexts, our system successfully got the recognition rate of 74.7%. We can also observe that without the context of speech or head movement the recognition accuracies decrease a lot. It means that these contexts play a significant role in detecting the interactions. On the other hand, interaction occasion and previous interaction are not as important as others because they do not have much influence on the detection result.

Table 2. Interaction detection results with different context sets (c1: head gesture, c2: attention from others, c3: speech tone, c4: speaking time, c5: interaction occasion, c6: type of previous interaction)

Table 1. Context extraction results

Context	Recognition rate	Context sets	Recognition rate
Head gesture	76.4%	Set 1 - all measures: {c1, c2, c3, c4, c5, c6}	74.7%
Attention from others	72.2%	Set 2 - all measures except head movement: {c3, c4, c5, c6}	66.8%
Speech tone	65.0%	Set 3 - all measures except speech feature: {c1, c2, c5, c6}	57.9%
		Set 4 - all measures except interaction occasion: {c1, c2, c3, c4, c6}	70.5%
		Set 5 - all measures except type of previous interaction: {c1, c2, c3, c4, c5}	72.6%

5. Conclusion

This paper describes an approach for collaborative capturing and detection of human interaction that is helpful for understanding social dynamics in meetings. As future work, we plan to integrate more contexts (e.g. lexical cues) in the detection process in order to improve the recognition accuracy. We also plan to design a visualization system for reviewing the human interactions.

6. Acknowledgements

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan under the projects of “Cyber Infrastructure for the Information-explosion Era”.

7. References

[1] Chang, C.C., and Lin, C.J., LIBSVM: a library for support vector machines, 2001. Software available at

<http://www.csie.ntu.edu.tw/~cjlin/libsvm>

- [2] Hsu, C.W., Chang, C.C., and Lin, C.J., A practical guide to support vector classification. Technical Report, 2005.
- [3] Nijholt, A., Rienks, R.J., Zwiers, J., and Reidsma, D., Online and Off-line Visualization of Meeting Information and Meeting Support, *The Visual Computer*, 2006, Vol. 22, No. 12, pp. 965-976.
- [4] Rabiner, L., A tutorial on Hidden Markov Models and selected applications in speech recognition, in: *Proc. IEEE*, vol.77, no.2, 1989, pp. 257-286.
- [5] Searle, J. *Speech Acts*, Cambridge University Press, 1969.
- [6] Stiefelhagen, R., Chen, X., and Yang, J., Capturing Interactions in Meetings with Omnidirectional Cameras, *International Journal of Distance Education Technologies*, 2005, Vol. 3, No.3, pp. 34-47.
- [7] Sumi, Y., et al, Collaborative capturing, interpreting, and sharing of experiences, *Personal and Ubiquitous Computing*, 2007, Vol. 11, No. 4, pp. 265-271.
- [8] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer Verlag, Heidelberg, DE, 1995.
- [9] Yu, Z., Ozeki, M., Fujii, Y., and Nakamura, Y., Towards Smart Meeting: Enabling Technologies and a Real-World Application, in: *Proceedings of ICMI'07*, pp. 86-93.