BSCHS (2008-13-2-9)
[Original Article]
Biomedical Soft Computing and Human Sciences, Vol.13, No.2, pp.67-74 (2008)

Event Based Dynamic Context Model for Group Interaction Analysis

Peng DAI, Linmi TAO and Guangyou XU

Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China

(received 9 October 2007, revised and accepted 29 November 2007)

Abstract: *Computer understanding of human social interactions is a challenging topic in the field of human computing due to its multi-party dynamic nature and multimodal characteristics. Context plays an essential role in the understanding of human behaviors during group interactions. This paper presents a novel Event Based Dynamic Context Model to represent hierarchical interaction context and solve the problems of context awareness. Sensing of dynamic context is based on multi-level event detection. Online analysis of multi-level events can be achieved by our model, which is superior over previous works. Implementations in our smart meeting room demonstrate the effectiveness of our approach.*

Keywords *Event Based Dynamic Context Model, Group interaction analysis, Meeting analysis, Event detection, Dynamic Bayesian Network.*

1. Introduction

The next generation computing will be about anticipatory user interfaces based on multiple intelligent sensors distributed in the environment, which should be human-centered and operate in the background [1]. The key research issue of human computing is that computer systems should analyze users' states and actions based on multimodal sensor data, and further provide attentive and non-intrusive services to users.

In our working and personal living spaces, social interactions with other people may occur frequently. Within the domain of human computing, computer understanding of human social interactions is an especially challenging topic due to its multi-party dynamic nature and multimodal characteristics. In this paper we address the problem of group interaction as the situation of multi-party face-to-face conversations, where group seminar meeting is a typical case.

Recently automatic analysis of group interactions in meeting scenarios has been a hot research topic. The EU research project M4 and its follow-up project AMI are the most well-known examples, which mainly dealt with group event analysis based on multimodal meeting corpus [2]-[4]. McCowan et al. [2] employed audio-visual information and applied Hidden Markov Models (HMM) for the classification of group actions in meeting scenarios. Zhang et al. [3] extended the work with a two-level HMM framework to model individual and group actions simultaneously. More recently new types of multimodal features such as prosody, speaker turns etc. were used and Multistream Dynamic Bayesian Model (DBN) was adopted to classify meeting scenarios [4]. However, most of the related works mentioned above were constrained to the offline mode, which means holistic analysis was performed based on the overall sequences. At the same time, previous works mostly classified meeting sequences into several basic types of overall scenarios, which could not denote variations of group interactions in different time scales and abstract levels.

However, our research objective is to make computers understand group situations online during meetings and provide attentive services accordingly. Therefore context awareness has to be achieved by computer systems. Context awareness plays a significant role in the domain of human centered computing, since context is tightly correlated to the analysis of human actions, interactions and intentions in two aspects. Firstly, appropriate understanding of human behavioral and social signals highly depends on the context, e.g. the same action may convey distinct meanings in different context. Secondly, context has to be considered for picking up the focus of attention in multimodal sensory data processing and fusion [1].

Context was defined previously to denote any information that can be used to characterize the situation that is relevant to the interaction between users and applications [5]. In this paper we separate environment and human related information in our context ontology definition and focus on human related context. Human related context ontology regarding group interactions can be structured as a hierarchy

Room 3-531, FIT Building, Tsinghua University,
Beijing 100084, China
Phone: +86-10-6278-2406
Fax: +86-10-6277-1138

according to different temporal scales and semantic abstract levels. The objective of this paper is to propose a general framework to appropriately represent and infer current context in group interaction scenarios.

Most of the previous works about context-aware computing applied context model to deal with the problems of context storage, sharing and management [6], which are not capable of handling the multi-level context sensing problem we faced in this paper. Recently some efforts have been made towards context-aware human activity analysis. Crowley [7] proposed a framework for context-aware observation of human activity. In [8] a distributed system was presented for best-perspective camera selection in multi-camera surveillance space, where dynamic context means user's locations and face orientations. However the context models in these literatures did not deal with multi-party group interactions. In [9] an event-driven context interpretation approach was presented to generate high-level contexts in Semantic Spaces. This literature mainly dealt with single-user situations and used logic inference for context reasoning, which cannot solve group interaction analysis problems based on multimodal sensory data fusion.

Due to the specialties of group interactions, we propose a novel Event Based Dynamic Context Model to represent contextual information and tackle context aware problems. A smart meeting room equipped with multimodal sensors is taken as our test bed. Within our model, context ontology hierarchy defines all the relevant contextual information in our applications. Multi-level events are defined in accordance with the context ontology hierarchy. Multi-level event detection forms the core of context aware engine. Lower level events are used as cues for higher level event inference, and reversely those inferred higher level events play the role of context in the guidance of lower level event detection. In our experiments, a flexible coarse-to-fine processing strategy and a refined probabilistic model are introduced respectively for online analysis. Our Dynamic Context Model has the advantage over previous methods in that it operates in an online mode and provides selectivity for other processing modules.

The rest of the paper is organized as follows. The scientific and practical reasons for investigating group interactions are explored in Section 2. Human related context in group interactions is introduced in Section 3. Section 4 describes our representation of context and operational mechanism of Dynamic Context Model. Implementations are presented in Section 5 and conclusions are drawn in Section 6.

2. Research Issues of Group Interactions

Automatic modeling of human group interactions from low-level multimodal signals is an interesting topic for both theoretical and practical reasons.

Firstly, group interactions can be recorded as multimodal sequences in the intelligent environment equipped with multiple sensors. Therefore, from the theoretical point of view, modeling of multi-party face-to-face conversations from multimodal sequences brings forward a particular challenging task for signal processing and machine learning techniques. Human social interactions have been investigated thoroughly in the field of social psychological research; however, computer modeling of human interactions still remains an open issue. Multi-party conversational group activities are categorized based on two axes in [10]. In the axis of temporal scale, group interactions can be classified from short term addressing and turn taking patterns to long term group trends and dominance. In the axis of group size, group interactions can be of dyadic, small and large groups. From our point of view, besides the factors concluded in [10], multimodal characteristic of human interactions also adds to the difficulty of group interaction modeling problem. During interactions, social signals may be conveyed by speech, focus of attention, and nonverbal actions such as iconic actions or hand actions accompanying speech.

Secondly, from the application point of view, automatic analysis and modeling of group interactions could add value to the unstructured raw data and make computers more helpful by providing attentive services to us such as online information delivery and retrieval.

By combining these two types of issues, we focus on group meeting analysis and present On-the-Spot Archiving System (OSAS) as our research platform, which is described in details in [11]. The main objective is to construct hierarchical semantic representation of group interactions from multimodal sensor data and provide attentive services meanwhile.

3. Human Related Context in Group Interactions

Although context awareness has been a hot topic in the field of ubiquitous computing, few related works have been conducted for group interactions, especially regarding human related context. Context is important for understanding of participants' activities within group interactions such as meetings. Previous works about group meeting analysis mostly dealt with offline classification problems [2], [3], [4], which means the problems are solved as information processing issues and no context awareness capabilities are considered.

Context awareness concerning group interactions has only been investigated in Computer Supported Cooperative Work (CSCW). Dourish and Bellotti [12] has presented early definition of awareness in CSCW systems: "*Awareness is an understanding of the activities of others, which provides a context for your own activity*". In group interactions, awareness of individual activities has to take the activities of others into account. However in [12] the users are distributed

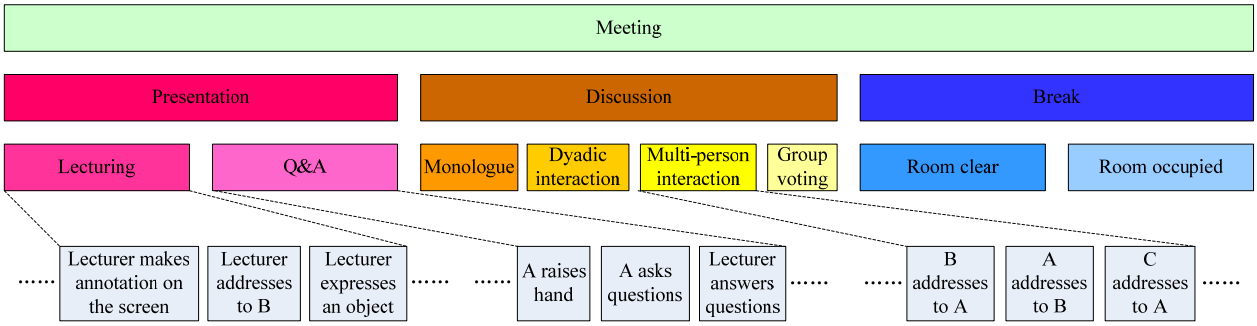


Fig.2. Human interaction context hierarchy in group meetings

and interact with each other through manual operations of computers, which are single-modal interactions. While in our case users talk to each other face-to-face and no explicit operations of computers are needed. Besides, multimodal characteristics of face-to-face interactions add to the complexity of our work.

Context in group interactions is not only individual activities, but also includes the entire group situation, mutual tasks and inter-personal relations. The relations between the group members are new factors we need to consider in the domain of context, which reside in two aspects: (1) Roles of each member in group interactions and in the task that the group is performing; (2) The structure of the group interactions.

Context in group interactions is dynamic, hierarchical, and high dimensional. During group interactions, roles may switch among different members and the interaction structure may change frequently. Therefore group interactions are dynamic spatial-temporal procedures, which makes context dynamic as well. Concerning different temporal scales and semantic abstract levels, group interactions can be decomposed as a hierarchical structure, which makes context be of hierarchy accordingly. Current context has to be inferred from the multimodal sensor signals, which makes the context sensitive to the high dimensional sensor data.

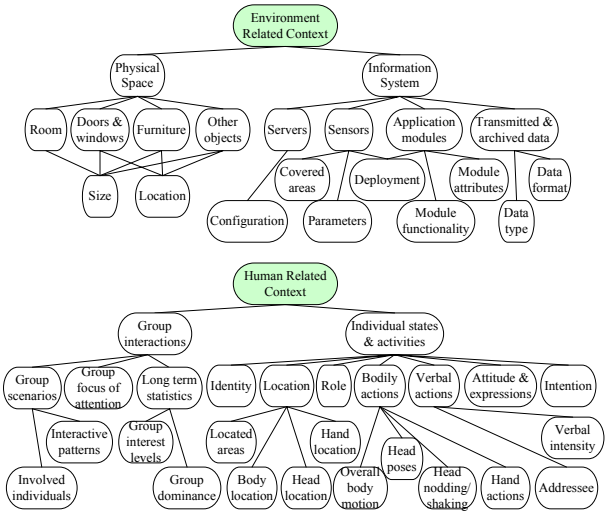


Fig.1. Context ontology toward group interactions

3.1. Context Ontology in Group Interactions

Greenberg defined context as "*a dynamic construct*" [13], which is in accordance with the dynamic nature of group interactions. The entire context of the group interaction environment should be composed of all the information related to human subjects, the physical and information environments [5]. In this paper, context ontology concerning group interactions is defined in Fig. 1, which can be concluded as two major categories: (1) environment related context, i.e. relevant information about the surrounding physical space and equipped information system; (2) human related context, which represents current and historical information about the states and activities of all participants.

In this paper we mainly focus on the human-human interaction context. As shown in Fig. 2, the interaction context within group meetings can be represented as a hierarchical structure in terms of temporal scales, which ranges from long term group scenarios, medium term interaction patterns to short term individual addressing switches. For instance, a "*group meeting*" may contain three types of group scenarios "*presentation*", "*discussion*" and "*break*" at long time scales. The "*discussion*" scenario may also contain four types of interaction patterns "*monologue*", "*dyadic interaction*", "*multi-person interaction*" and "*group voting*". In shorter time scales, the sub-scenario "*multi-person interaction*" may be segmented into various stages, such as "*A addressing to B*", "*C addressing to A*" etc.

4. Event Based Dynamic Context Model

Semantic analysis results of group interactions based on multimodal sensor data play the role of context, which can be used to guide the sensor signal processing and semantic analysis processes. However how to make computers aware of current context still remains to be a challenging task. In this paper, the research issue of context awareness is concluded as effective integration of bottom-up context reasoning and top-down context guidance in a consistent framework.

In group interactions, human related context is dynamic in both the spatial and temporal dimensions, and it is comprised of multiple semantic abstract levels.

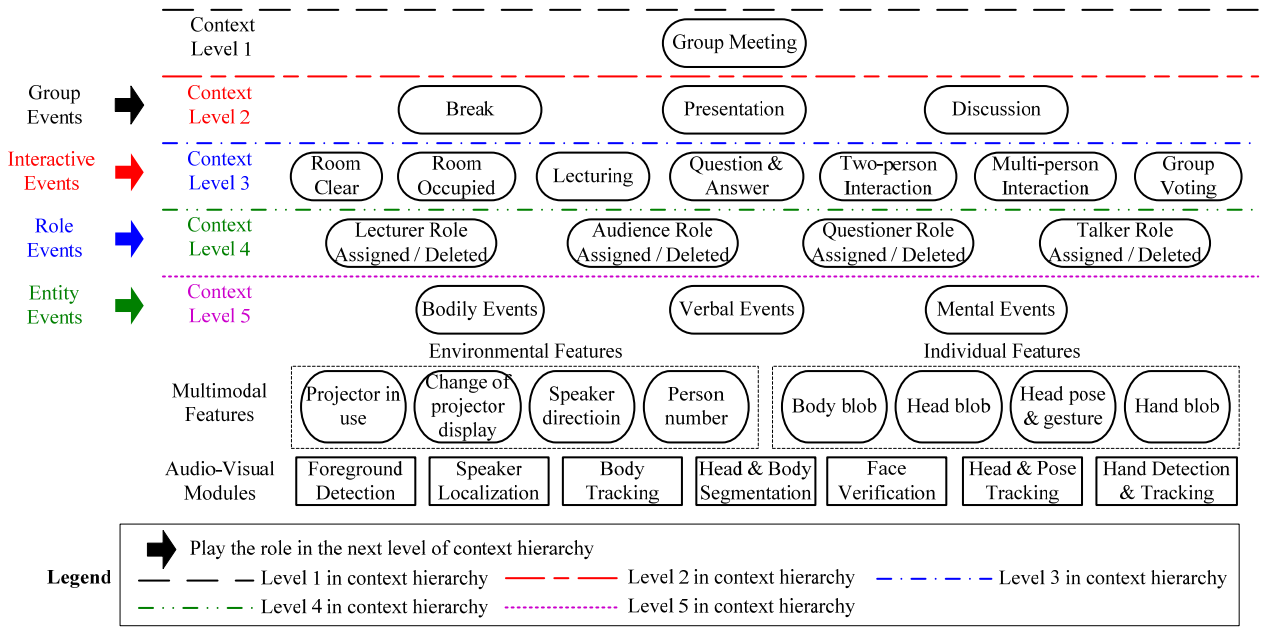


Fig.3. Relationship between context and event hierarchy

Thus current context cannot be determined simply by collecting low level sensor data as in most of the previous context aware systems [6]; on the contrary it has to be inferred based on multi-level event detection mechanism, which conveys fundamental context parameters of *who*, *when*, *where*, *what*, *why* and *how* (5W1H) at different abstract levels. An event based model for the representation and reasoning of dynamic context is presented in this paper, which serves as the basis of context awareness in group interactions.

4.1. Context and Events

Context and events are closely related to each other, i.e. events are defined under specific context, and context is inferred through detecting events by the computer system. Crowley defined events to represent changes in situation that can be used to trigger system actions [7]. In this paper we define two types of events: "*switching event*" that results in situation switches, and "*characteristic event*" that characterizes current situation and does not trigger situation changes. Events in group interactions fall into four abstract levels: group, interactive, role and entity level, which is in correspondence with the hierarchical structure of human-human interaction context. At the group level, three typical meeting scenarios "*presentation*", "*discussion*" and "*break*" are defined as group events.

The relationship between context hierarchy and event hierarchy is illustrated in Fig. 3. Four context levels are expressed according to various temporal scales. At context level 1, given the "*group meeting*" scenario, three types of group events can be detected, which further play the role of context at context level 2. At the bottom level, detected entity events play the role of context at context level 4, which can be applied as guidance for the selectivity of audio-visual processing modules and fusion of multimodal features.

4.2. Dynamic Context Model

In this paper, a novel Event Based Dynamic Context Model is proposed for context representation and reasoning. The conceptual model is defined as

$$M = \{O, S, E, R, F\}, \quad (1)$$

among which O denotes context ontology, S represents current situation, E denotes event hierarchy, R expresses the relationship between context and event hierarchy, and F represents current set of multimodal features.

As has been defined previously, context ontology is categorized as two major types: environment context ontology O_E, and human context ontology O_U. Current situation S can be instantiation of context ontology at current time slice, which can also be concluded in two categories S_E and S_U corresponding to current environment situation and human interaction situation respectively. Current interaction situation is organized in a hierarchical structure $S_U = \{S_U^A, S_U^G, S_U^V, S_U^R, S_U^E\}$, which represents situation at overall scenario level (i.e. "*group meeting*"), group level, interactive level, role level and entity level, as shown in Fig. 3.

Human related context, i.e. interaction context in this paper, serves as the key of context awareness toward group interaction analysis. Based on the definition of context and event hierarchy in Section 4.1, online detection of multi-level events constitutes the core of context reasoning. Thus we define an event hierarchy $E = \{E_G, E_V, E_R, E_E\}$, which is in accordance with the current human situation at group, interactive, role and entity levels, as is expressed in Fig. 3.

Denote interaction situation and event hierarchy as $S_U = \{S_U^i, i = 1, \cdots, 5\}$ and $E = \{E_i, i = 1, \cdots, 4\}$, the relationship between human interaction situation and

multi-level events can be represented as:

$$\begin{aligned} &\{F, E_j, j=1,\cdots,i-1\} \overset{Reasoning}{\Rightarrow} E_i \\ &\{E_j, j=1,\cdots,i-1\} \overset{Switching}{\Rightarrow} S_U^i \\ &S_U^i \overset{Guidance}{\Rightarrow} \{proc(F), proc(E_j), j=1,\cdots,i\} \end{aligned} \quad , \qquad (2)$$

which concludes the operational mechanism of our model into three parts: higher level events are detected based on lower level events and multimodal features, detected events can be converted to contextual information at the corresponding level, and context at each level can be applied as guidance for the procedures of lower level event detection and multimodal signal processing. Such relationship between multi-level events and multi-level situation is illustrated in Fig. 4.

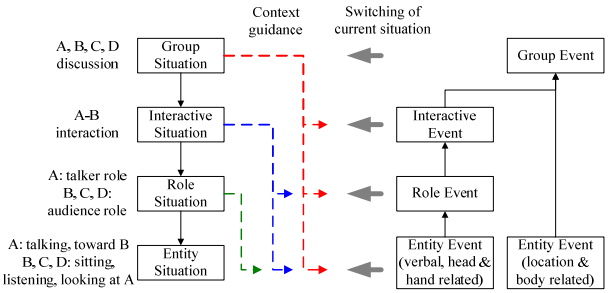


Fig.4. Relationship between situation and events

The essence of our Dynamic Context Model is to sense current context based on multimodal sensor data, and further provide context-aware services to the users automatically. Thus contextual information definition and organization, context sensing engine, and context guiding engine are presented as the three major components of our model, which are tightly integrated with each other in the structure illustrated in Fig. 5.

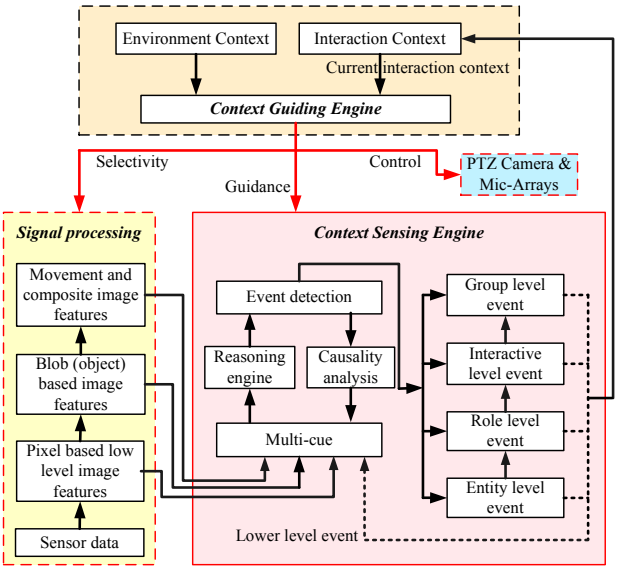


Fig.5. Structure of Dynamic Context Model

Context sensing engine holds the key to the operation of our model. At any time step, bottom-up event reasoning and top-down context guidance can be performed online. Our model takes an effective operating mode to solve the problem of context sensing, which can be concluded as two main characteristics: (1) bottom-up and top-down integrated event detection, (2) coarse-to-fine feature detection. When the computer system gets started, current situation is unknown and we only detect human presence and some global features, based on which coarse level judgment about current group situation can be achieved. Then it can be further applied as contextual guidance to detect refined features and more events at detailed levels. The main functionality of contextual guidance includes: (1) selectivity among various audio-visual modules so as to reduce computational cost, (2) determining which events to be detected and monitored in current context, (3) generating control commands for the active sensors such as Pan-Tilt-Zoom camera and microphone arrays.

5. Implementations

We have applied the conceptual model to the OSAS system in our smart meeting room, which are equipped with multiple sensors including three fixed cameras, three linear microphone arrays and one Pan-Tilt-Zoom camera. Our signal processing modules, context reasoning modules, archiving and retrieving modules are all designed based on the extracted audio and visual information.

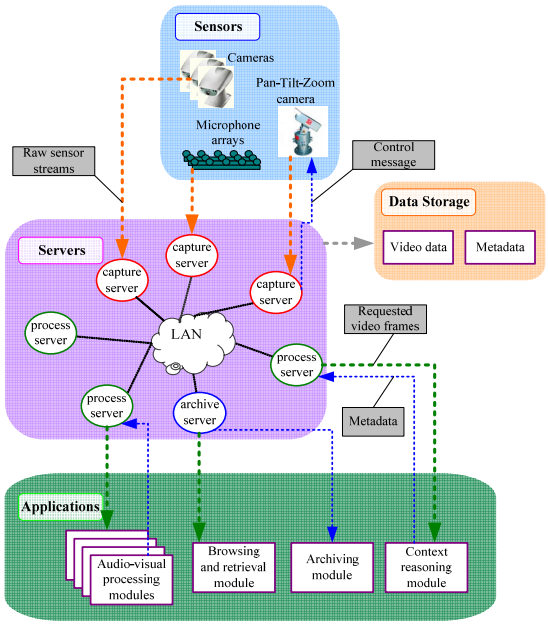


Fig.6. Implemented system framework

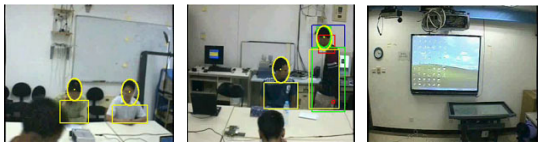
A flexible multi-server platform is adopted as the software infrastructure to support distributed multimedia processing, which includes such functions as sensor data capture, transmission, analysis, storage and retrieval [15]. The distributed framework makes our

system more efficient and robust. Multiple processing modules are distributed on various servers and can operate simultaneously. They communicate with each other through LAN, and transmitted data can be divided into three types: raw audio-visual streams, MPEG7-format metadata, and control command. Fig. 6 gives an overview of the software system.

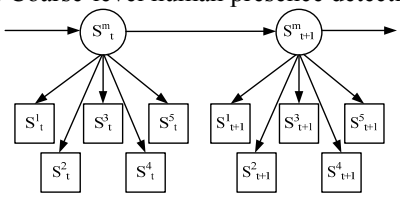
Based on the aforementioned platform, two types of implementations have been brought forward to evaluate the flexibility and effectiveness of our conceptual context model.

5.1. Context-Aware Signal Processing Mechanism

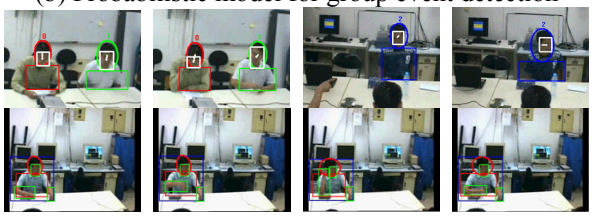
An effective multimodal information processing mechanism needs to be defined so as to abandon those unnecessary processing modules and reduce computational cost according to current context, especially in an online processing system like ours. Therefore an effective context-aware signal processing mechanism is proposed in our first implementation. Only two levels of the interaction context hierarchy are involved, which adopts a coarse-to-fine strategy and solves the problem of event detection in two different probabilistic models [14]. Group level events are detected through a single layer Dynamic Bayesian Network based on coarse global features such as human locations. If current group scenario is classified as "*presentation*" or "*discussion*", refined level analysis will be performed at the entity level, such as detecting entity verbal events through Bayesian Network based on microphone array signals, detecting entity pose events and hand events based on visual signals. Although the implementation only deals with two levels of interaction context, such kind of coarse-to-fine online processing strategy applies context awareness in the guidance of multimodal signal processing tasks, especially vision tasks, and reduces the computational cost effectively. The overall workflow of our mechanism is shown as follows:



(a) Coarse-level human presence detection



(b) Probabilistic model for group event detection



(c) Refined-level processing, e.g. pose estimation and hand detection

Fig.7. Context-aware signal processing

5.2. Multi-level Probabilistic Reasoning Model

As the computer system comes to a stable stage, our Dynamic Context Model needs to detect multi-level events simultaneously so as to understand the overall situation. By integrating multi-level event detection in a unified reasoning model, we can reduce random errors in the separate reasoning of events at each level. Therefore in our second implementation, we present a novel probabilistic model named Event Driven Multi-level Dynamic Bayesian Network (EDM-DBN) to model hierarchical interaction context and perform online analysis of multi-level events, as shown in Fig. 8. The EDM-DBN model analyzes semantic information at more refined levels, which is in accordance with the context and event hierarchy defined in Fig. 3. Our approach integrates the bottom-up reasoning and top-down guidance together to form a consistent reasoning framework, which has the advantage over previous methods in that it detects multi-level events simultaneously online. More inference details about our EDM-DBN model have been introduced in [11].

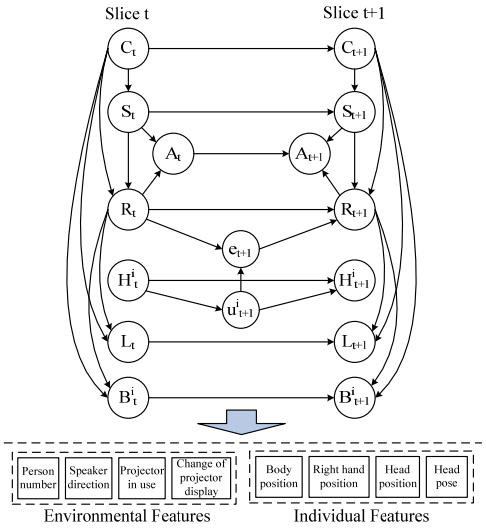
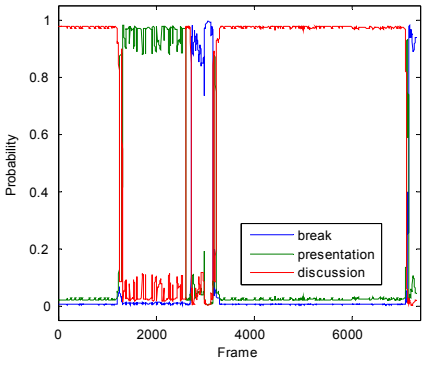


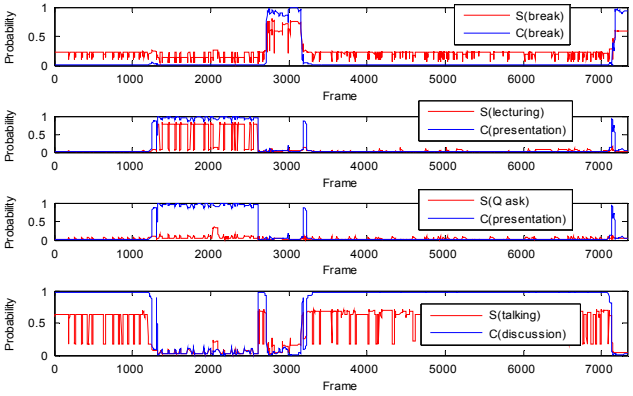
Fig.8. Our multi-level probabilistic reasoning model

Some experimental results are shown in Fig. 9. Here we only give the output results of two high-level semantic nodes C_t and S_t, which corresponds to group level and interactive level situation respectively. From the results we can see that our multi-level probabilistic model can not only distinguish the overall group situations "*break*", "*presentation*" and "*discussion*", as shown in Fig. 9(a), but also it can identify those interactive situations under certain group scenarios. For instance, our model can classify "*lecturing*" and "*question & answer*" sections under "*presentation*" scenarios, as illustrated in Fig. 9(b). And it can distinguish various speaking sequences of participants

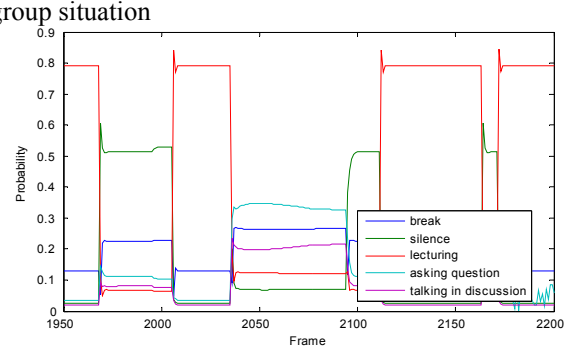
in "*discussion*" situations by taking both S_t and R_t (i.e. role of each participant) into consideration.



(a) Output result of node C_t, i.e. probabilistic result of group situation



(b) Output result of node S_t, i.e. probabilistic result of interactive situation regarding the current corresponding group situation



(c) Segmented output result of node S_t

Fig.9. Experimental results with EDM-DBN

Experimental results of the implementations demonstrate the effectiveness of our framework. Although context reasoning results with our probabilistic models are uncertain values, we can choose the values with maximum probabilities as the output results of context at each level, which can be stored in MPEG7 based meeting archives and used for future retrievals.

6. Conclusion

Context has to be considered while analyzing human actions and interactions in the domain of human computing. Sensing contextual information in group interactions is especially challenging due to the dynamic nature, hierarchical characteristic and high dimensionality of interaction context. This paper presents a novel Event Based Dynamic Context Model for the solution of context-aware computing in meeting scenarios. The essence of context sensing is multi-level event detection, which integrates bottom-up reasoning and top-down guidance together consistently. In our experiments, a flexible coarse-to-fine signal processing strategy and a refined probabilistic model are implemented respectively based on the conceptual model. Online analysis and selectivity among multiple modules is characterized as a great advantage of our approach over previous works. The effectiveness of the proposed framework has been tested in our smart meeting room environment. Future work includes improvement of reasoning model and integration of proactive services.

Acknowledgments

The work described in this paper was supported by National Science Foundation of China (No. 60673189) and National Science Foundation of China (No. 60433030).

References

[1] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2006) "Human Computing and Machine Understanding of Human Behavior," A Survey, Proc. 8th Intl. Conf. Multimodal Interfaces, pp.239-248.

[2] McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D. (2005) "Automatic Analysis of Multimodal Group Actions in Meetings", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.27, No.3, pp.305-317.

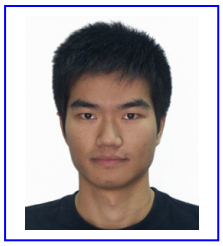
[3] Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I. (2006) "Modeling individual and group actions in meetings with layered HMMs," IEEE Trans. Multimedia, Vol.8, No.3, pp.509-520.

[4] Dielmann, A., and Renals, S. (2007) "Automatic Meeting Segmentation Using Dynamic Bayesian Networks," IEEE Transactions Multimedia, Vol.9, No.1, pp.25-36.

[5] Dey, A.K., Abowd, G.D., and Salber, D. (2001) "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications," Journal of Human-Computer Interaction, Vol.16, pp.97-166.

[6] Baldauf, M., Dustdar, S., and Rosenberg, F. (2006) "A Survey on Context-Aware Systems," Intl. Journal of Ad Hoc and Ubiquitous Computing.

[7] Crowley, J.L. (2003) "Context Driven Observation of Human Activity," Proc. European Symposium on Ambient Intelligence, pp.101-118.
[8] Trivedi, M.M., Huang, K.S., and Mikic, I. (2005) "Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces," IEEE Trans. Systems, Man, and Cybernetics—PART A: Systems and Humans, Vol.35, No.1, pp.145-163.
[9] Tan, J.G., Zhang, D., Wang, X., and Cheng, H.S. (2005) "Enhancing Semantic Spaces with Event-Driven Context Interpretation," Proc. Intl. Conf. Pervasive Computing, pp.80-97.
[10] Gatica-Perez, D. (2006) "Analyzing group interactions in conversations: A review," Proc. IEEE Intl. Conf. Multisensor Fusion and Integration for Intelligent Systems, pp.41-46.
[11] Dai, P., and Xu, G. (2007) "Event Driven Dynamic Context Model for Group Interaction Analysis," Intl. Conf. Soft Computing and Human Sciences.
[12] Dourish, P., and Bellotti, V. (1992) "Awareness and Coordination in Shared Workspaces," Proc. ACM Conference on Computer Supported Cooperative Work, Toronto, Ontario, ACM Press.
[13] Greenberg, S. (2001) "Context as a dynamic construct," Journal of Human-Computer Interaction, Vol.16, pp.257-268.
[14] Dai, P., Tao, L., and Xu, G. (2007) "Audio-Visual Fused Online Context Analysis toward Smart Meeting Room," Proc. 4th Intl. Conf. Ubiquitous Intelligence and Computing, pp.868-877.
[15] Wang, Y., Tao, L., Liu, Q., Zhao, Y., and Xu, G. (2007) "A Flexible Multi-Server Platform for Distributed Video Information Processing," Proc. 5th Intl. Conf. Computer Vision Systems, Germany.



Peng DAI
He received the B.E. degree in computer science from Tsinghua University, Beijing, China, in 2002. He is currently a Ph.D. candidate in Tsinghua National Laboratory for Information Science and Technology, Tsinghua University. His research interests include computer vision and multimodal human computer interaction.

Linmi TAO
He is an associate professor at Dept. of Computer Science, Tsinghua University. He received his B.S. on Biology, M.S. in cognitive science and Ph.D. degree on computer science. He has studied and worked in Italy in International Institute for Advanced Scientific Studies and University of Verona on computational visual perception, 3D visual information processing. His researches cover a brand spectrum on computer vision, cognitive vision, affective computing, and bioinformatics.

Guangyou XU
He received B.E. degree with distinction in Automatic Control Engineering from Tsinghua University in 1963. He was a visiting scholar at School of Electronic Engineering, Purdue University, U.S. from 1982 to 1984. He served as Director of Information Processing and Application Division at Dept. of Computer Science, Tsinghua University from 1986 to 1997. He was a Visiting Professor at Beckman Institute, University of Illinois at Urbana-Champaign from 1993 to 1994. His current research interests include computer vision, human computer interaction, and pervasive computing. He is a member of Robot Measurement Committee TC-17 of IMEKO, a standing member of the Council of China Image and Graphic Association and Chair of Multimedia Technology Committee, China Image and Graphic Association. He is the Associate Editor of International Journal of Biomedical Soft Computing and Human Sciences , the Associate Editor of International Journal of Image and Graphics.