Event Driven Dynamic Context Model for Group Interaction Analysis

Peng Dai
Tsinghua National Lab on Information Science & Technology
Tsinghua University
Beijing, China
daip02@mails.tsinghua.edu.cn

Guangyou Xu
Tsinghua National Lab on Information Science & Technology
Tsinghua University
Beijing, China
xgy-dcs@tsinghua.edu.cn

Abstract: Computer understanding of human actions and interactions is the key research issue in human computing. Meanwhile context has been considered to play an essential role in understanding of human behavior during group interactions. This paper proposes a novel Event Driven Dynamic Context Model underlying the context sensing framework to support online analysis of group interactions in meeting scenarios. Sensing of dynamic context is based on multi-level event detection. A novel probabilistic model named Event Driven Multi-level Dynamic Bayesian Network (EDM-DBN) is presented to model hierarchical context and perform online analysis of multi-level events, which is superior over previous works. Experimental results in our smart meeting room demonstrate the effectiveness of our approach.

Keywords: Dynamic Context Model, event driven, group interaction analysis, multimodal meeting analysis.

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' actions and intentions based on multi-modal sensor data, and further provide proactive services which are non-intrusive to users. Context awareness plays a significant role in the domain of human 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 process and fusion of sensory data across multiple modalities [1].

In this paper, the problem of context awareness is defined as online analysis of human actions, intentions and overall situation toward multi-party face-to-face conversations such as group meetings in a meeting room. Dourish and Bellotti [2] has presented early definition of awareness in Computer Supported Cooperative Work (CSCW) systems: "*Awareness is an understanding of the activities of others, which provides a context for your own activity*". In our work, awareness of individual activities and intentions also take the activities of others into account. However in [2] the group users are distributed and interact with each other through manual operations of computers; while in our work users interact with each other face-to-face and no explicit operations of computers are needed in our system. Besides, requirements of multimodal information fusion add to the complexity of our work.

Analysis result of individual activities, intentions and group situations as a whole can be regarded as context, which is dynamic and hierarchical in group interaction scenarios. Greenberg defined context as "*a dynamic construct*" [3], which is in accordance with the dynamic nature of group interactions in meetings. The entire context of the group interaction environment should be composed of all the information related to human subjects, the physical and information environments. For the sake of simplicity in analysis, we divided context into two parts: "*environment context*" that is related to the physical and information environment, and "*interaction context*" that concerns the interactive situations among people. As is shown in Fig. 1, the "*interaction context*" of meetings is of a hierarchy in terms of temporal scales, which will be sensed by means of detecting events at each hierarchical level correspondingly. In the mean time events detected at higher level will play the role of context to guide the detection of lower level event.

Most of the previous works on context-aware computing presented context model to deal with the problems such as context storage, sharing and management in the field of ubiquitous computing [4], which are not capable of handling the multi-level context sensing problem we face in this paper. Recently some efforts have been made towards the context-aware visual computing in smart environment.

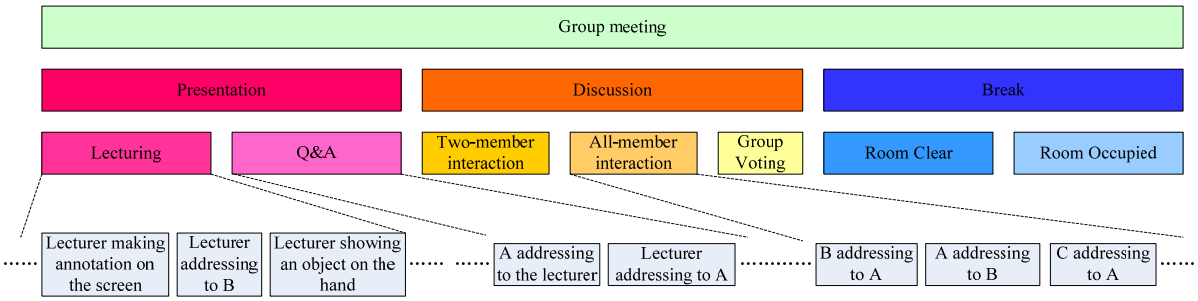

Fig. 1. Interaction context hierarchy in meetings

Crowley [5] proposed a framework for context-aware observation of human activity. Trivedi et al [6] presented a distributed system paradigm for dynamic context capture in indoor environment, where dynamic context means changes of user's location and face orientation. However the context models in these literatures did not deal with such composite group interactions which consist of multiple abstract levels as in this paper. In [7] an event-driven context interpretation approach was presented to generate high-level contexts in Semantic Spaces. This literature focused on ubiquitous computing problems, mainly dealt with single-user situations and used logic inference for event and context reasoning. However this method cannot solve the multi-party group interaction analysis problems, which involves multi-user interaction situations and the hierarchical events detection based on multimodal sensor information fusion. In this regard, a probabilistic reasoning model is required to detect events at multiple context levels as we did in this paper.

Therefore we propose an event driven approach to solve the context awareness problem toward group interactions in meetings. There have been some related works on automatic recognition of individual or group events in meeting scenarios. The EU research project M4 and its follow-up project AMI mainly dealt with group event analysis based on multimodal meeting corpus [8], [9], [10]. McCowan et al. [8] employed both the audio and visual information and applied Hidden Markov Models (HMM) for the recognition of group actions in meeting scenarios. Zhang et al. [9] 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 to classify meeting scenarios, and Multistream Dynamic Bayesian Model (DBN) was adopted in [10]. 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 only after meetings had finished.

In this paper, a novel Event Driven Dynamic Context Model is proposed to tackle context aware problems for group interactions. Multi-level events are defined corresponding to the context hierarchy. Lower level events and bottom level sensor data can be used as cues for higher level event inference. Those inferred higher level events play the role of context to guide the detection of lower level events. A novel probabilistic model named Event Driven Multi-level Dynamic Bayesian Network (EDM-DBN) is presented to detect the multi-level events online, which integrates the bottom-up reasoning and top-down guidance together to form a consistent reasoning framework. Our approach has the advantage over previous methods in that it detects multi-level events simultaneously online.

The rest of the paper is organized as follows. Event Driven Dynamic Context Model is introduced in Section 2. A novel probabilistic model is described in Section 3 to detect multi-level events. Experimental results are presented in Section 4 and conclusions are drawn in Section 5.

2. Event Driven Dynamic Context Model

Understanding of group interactions has to incorporate contextual information. However how to make computers aware of current context still remains a challenging task for context-aware systems. In group interactions, context is dynamic in both the spatial and temporal dimensions, thus it cannot be determined simply by collecting low level sensor data; on the contrary, context has to be inferred based on detected events, which convey semantic meanings at different abstract levels. This paper aims to provide an event driven probabilistic framework for the representation and inference of dynamic context, which serves as the basis of context-ware computing system.

2.1. Research Platform

Our objective is to make computer systems understanding group interactions and provide proactive services during group seminar meetings, such as active control of Pan-Tilt-Zoom (PTZ) camera and microphone arrays, and automatic online archiving of meeting data with multi-level semantic information. Thus an intelligent On-the-Spot Archiving System (OSAS) has been developed as our

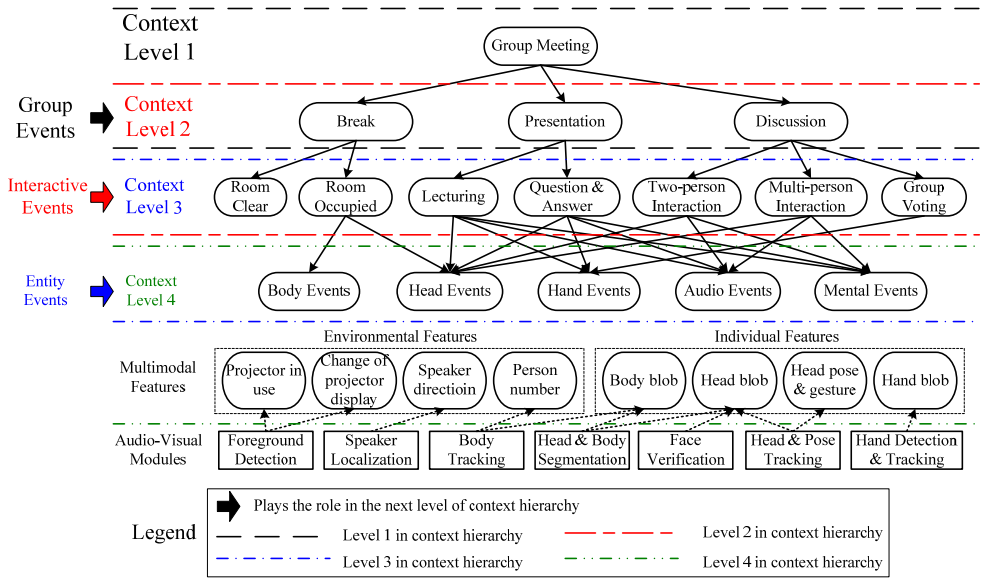

Fig. 3. Context and event hierarchy

research platform.

The meeting room environment setting for the project is shown in Fig. 2. Three fixed cameras are deployed to monitor the meeting room from distinct perspectives. A Pan-Tilt-Zoom (PTZ) camera is placed on the table to focus on any specific target. Three linear microphone arrays are installed on the table to collect speech information from various participants. Besides, a flexible multi-server platform is adopted as the software infrastructure to support distributed multimedia information processing.

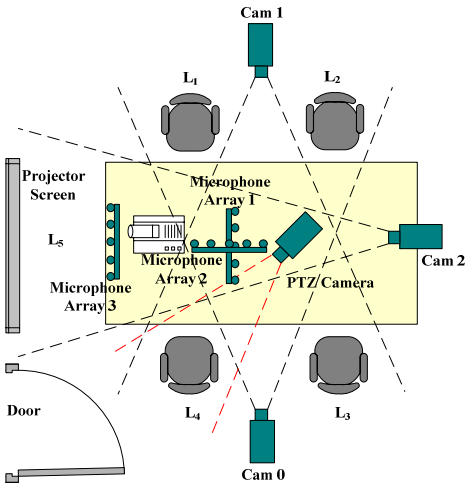

Fig. 2. Meeting room setting

2.2. Context and Events

In this paper we mainly focus on the "*interaction context*". As is shown in Fig. 1, "*interaction context*" of group meetings can be divided into multiple layers according to various temporal scales. For instance, a "*group meeting*" may contain three types of group situations "*presentation*", "*discussion*" and "*break*" at smaller time scales. The "discussion" situation may also contain three types of sub-situations "*two-member interaction*", "*all-member interaction*" and "*group voting*". In the sub-situation "*all-member interaction*", the discussion procedure might contain various stages, such as "*A addressing to B*", "*C addressing to A*" etc.

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 [5]. 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 correspond to the hierarchical structure of "*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 and are applied to guide selectivity and fusion of multimodal features and audio-visual processing modules.

2.3. Dynamic Context Model

Based on the definition of context and event hierarchy, online detection of multi-level events constitutes the core of context aware engine. As is

illustrated in Fig. 4, higher level events are detected through a Context Sensing Engine based on lower level events and multimodal features. At the same time, higher level events will serve as contextual information for the guidance of lower level event detection, which is achieved through the Context Guiding Engine. The multimodal features used in multi-level event detection are extracted by various signal processing modules.

Once the multi-level events are detected, they are taken as the contextual information by the Context Guiding Engine, which then performs top-down guidance based on the current context. Its main functionality includes selectivity and fusion among various audio-visual modules, determining which actions to be monitored in current situation, and generating control commands for the active sensors.

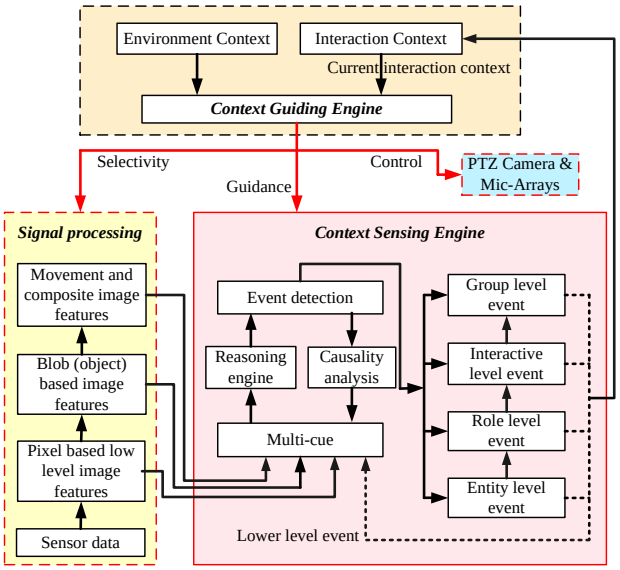

Fig. 4. Structure of Dynamic Context Model

The structure of our Dynamic Context Model is illustrated in Fig. 4, within which current context, Context Sensing engine and Context Guiding Engine are tightly integrated. 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 needs to be performed online, which is guaranteed by the probabilistic reasoning framework presented in Section 3.

3. Probabilistic Model for Multi-Level Event Detection

According to the previous section, the overall context is sensed by detecting and integrating events at all hierarchical levels. Lower level events and multimodal features are used as multiple cues for the inference of higher level events. Higher level events play the role of context in detecting lower level events. In this section, a novel probabilistic model is presented to model the relationships between multi-level events and multimodal features.

3.1. Event Driven Multi-level Dynamic Bayesian Network

A novel probabilistic model named Event Driven Multi-level Dynamic Bayesian Network (EDM-DBN) is proposed in this paper to model dynamic context toward group interactions, as is shown in Fig. 5. The state nodes are defined corresponding to the multi-level events. Given the meeting scenarios, node C_t at the top level represents current group events during meetings. It contains several types of interactive events at the second level node S_t. For instance, "*presentation*" may consist of "*lecturing*", "*silence*", and "*question asking*". S_t happens in the context of C_t, i.e. state transitions of node S_t depend on its upper-level event C_t.

We model the scenarios in a location-aware manner, i.e. we are interested in what happened at some specific locations. In this paper, five interest locations *A*, *B*, *C*, *D*, and *P* are defined, which represent the four seat locations around the meeting table and the location of projector screen respectively. Only when participants are at these locations, their actions can be regarded meaningful for group interaction analysis.

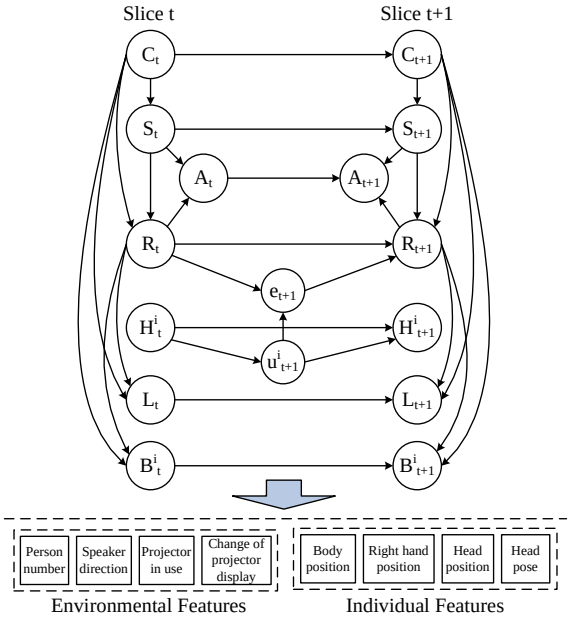

Fig. 5. Probabilistic model for multi-level event detection

Except for the node C_t and S_t, the other state nodes in our model are all defined related to the users' locations inside the meeting room. A set of location-aware nodes L_t^G $(G = \{A, B, C, D, P\})$ are defined to describe whether any participant appears at any one of the interest locations. Two other

nodes B_t^i and H_t^i $(i=\{1,2,3,4\})$ are defined to characterize what is the body status and hand status of each participant, such as sitting or standing, hand on table, put up or holding together. It can be observed that most sub-scenarios of meetings are of speech interactions, thus audio is a helpful channel for the analysis. Audio state node A_t denotes the speaker identification during meetings, i.e. from which location the speech comes. Role node R_t is introduced to simplify the modeling logic, which includes three different types of roles R_t^G $(G=\{P,Q,T\})$. Among them, R_t^P indicates seat index from which the lecturer comes and R_t^Q denotes the location from which the question comes during the "*presentation*" situation, while R_t^T indicates which seat the speaker takes during "*discussion*". For simplification, L_t^G and R_t^G are grouped in Fig. 5, and only entity event nodes from one participant are displayed.

Besides state nodes described above, several policy nodes are also defined that can be applied to trigger the temporal switching of some state nodes. For instance, node e_{t+1} represents switching policies of role nodes, node u_{t+1}^i denotes policies about transitions of hand states concerning each participant. States of C_t, B_t^i, H_t^i and A_t are related to the individual and environmental features, however they are not expressed in Fig. 5 for clearness.

We take the audio state node A_t and location-aware node L_t^A as an instance to explain causal relations and state transitions in our probabilistic model. The likelihood function of speech state node A_t is defined as:

$$P(\alpha_t \mid A_t = a) = N(\cdot;\beta_a,\theta^2) \tag{1}$$

where α_t denotes speaker angles extracted from microphone arrays, β_a is the mean angle of direction *A*. The likelihood function of location-aware node L_t^A is written as:

$$P(Y_t^i \mid L_t^A = b) = N(Y_t^i(x); x_A, \sigma^2) \tag{2}$$

where Y_t^i indicates i-th participant's head position, and x_A is the mean of *x* coordinate at interest location *A*. Probabilistic transitions of audio states are denoted as follows:

$$\begin{aligned} P(A_{t+1}^i = b \,\big|\, A_t^i = a, S_{t+1} = s, R_{t+1}^Q = q, R_{t+1}^T = t) \\ = \pi_A^s(a,b) \end{aligned} \tag{3}$$

in which audio state A_{t+1} is reasoned from the combination of previous state A_t, current interactive state S_{t+1}, questioner role R_{t+1}^Q and talker role R_{t+1}^T. Interactive event is taken as contextual information to guide the inference of audio events in a top-down mode, i.e. audio state transition is adjusted according to different upper-level context S_{t+1}. State transitions of node L_t, R_t, and B_t^i can be generated in similar ways.

The proposed approach expresses the causal relations between multi-level events and low level multimodal features. Bottom-up driven and top-down guidance has been blended in the model and solves the problem of context awareness successfully.

3.2. Model Inference Approach

The inference method of the proposed EDM-DBN model is introduced in this sub-section. Online analysis of group interactions can be summarized as the problem of making inference at each time step and generating output results of state nodes at all levels simultaneously.

According to the model structure, at time step *t+1* we have two sets of hidden variables: state nodes:

$$h_{t+1}^s \equiv \begin{Bmatrix} C_{t+1}, S_{t+1}, R_{t+1}^{P,Q,T}, L_{t+1}^{A,B,C,D,P}, A_{t+1}, \\ B_{t+1}^i, H_{t+1}^i, i=1,2,3,4 \end{Bmatrix} \tag{4}$$

and policy nodes:

$$h_{t+1}^e \equiv \{e_{t+1}, u_{t+1}^i, i=1,2,3,4\} \tag{5}$$

Observations of our model are multimodal features extracted with multiple audio-visual processing modules, which contains audio features recorded by microphone arrays, visual features about projector screen area, participants' hand blobs, head blobs and head poses. Given the observations, our objective is to calculate the posterior probability as follows:

$$p(h_{t+1}^s, h_{t+1}^e \mid O_{0:t+1}) \propto p(h_{t+1}^s, h_{t+1}^e, O_{0:t+1}) \tag{6}$$

which can be further represented with the following equation:

$$\begin{aligned} &p(h_{t+1}^s, h_{t+1}^e, O_{0:t+1}) = p(O_{t+1} \mid h_{t+1}^s) \cdot \\ &\sum_{h_t^s} p(h_{t+1}^e \mid h_t^s, O_{0:t}) p(h_{t+1}^s \mid h_t^s, h_{t+1}^e) p(h_t^s \mid O_{0:t}) \end{aligned} \tag{7}$$

As $p(h_{t+1}^e \mid h_t^s, O_{0:t})$ is modeled directly by the EDM-DBN, we can just calculate it, and use the results of h_{t+1}^e to drive and optimize h_{t+1}^s. An efficient joint inference algorithm based on dynamic programming is proposed in this paper to obtain global optimization of h_{t+1}^s given $O_{0:t+1}$. The inference algorithm is listed in Table 1 below.

Table 1. Inference algorithm for EDM-DBN

1) Elimination
Calculate $p(A_{t+1} \mid O_{0:t+1}, S_{t+1}, R_{t+1}^Q, R_{t+1}^T)$, and sum A_{t+1} out.
Calculate $p(R_{t+1}^T \mid O_{0:t+1}, S_{t+1}, R_{t+1}^Q)$, and sum R_{t+1}^T out.
Calculate $p(R_{t+1}^Q \mid O_{0:t+1}, C_{t+1}, S_{t+1})$, and sum R_{t+1}^Q out.
Calculate $p(B_{t+1}^i \mid O_{0:t+1}, C_{t+1}, R_{t+1}^P)$, and sum B_{t+1}^i out.
Calculate $p(L_{t+1}^{A\sim D} \mid O_{0:t+1}, C_{t+1}, R_{t+1}^P)$, and sum $L_{t+1}^{A\sim D}$ out.
Calculate $p(S_{t+1} \mid O_{0:t+1}, C_{t+1})$, and sum S_{t+1} out.

Calculate $p(R^P_{t+1} | O_{0:t+1}, C_{t+1})$, and sum R^P_{t+1} out.
Calculate $p(L^P_{t+1} | O_{0:t+1}, C_{t+1})$, and sum L^P_{t+1} out.
Finally, obtain $p(C_{t+1} | O_{0:t+1})$.
2) Recursive back
Calculate $p(S_{t+1} | O_{0:t+1})$, $p(L^P_{t+1} | O_{0:t+1})$, $p(R^P_{t+1} | O_{0:t+1})$, $p(L^{A\sim D}_{t+1} | O_{0:t+1})$, $p(B^i_{t+1} | O_{0:t+1})$, $p(R^Q_{t+1} | O_{0:t+1})$, $p(R^T_{t+1} | O_{0:t+1})$, and $p(A_{t+1} | O_{0:t+1})$ respectively.

Based on the model inference approach introduced above, output results of all the state nodes in our EDM-DBN model can be generated simultaneously at each time step. Our approach ensures the online detection of multi-level events in group interactions.

4. Experiments and Discussion

Our approach has been tested in the smart meeting room environment. Video sequences are extracted from three fixed cameras, and audio signals are generated from three microphone arrays. Each test meeting lasts about 5 minutes. Audio signals and video data are synchronized through time stamps. Fig. 6 shows some sample video frames used in our experiments.

Fig. 6. Sample video frames extracted and used in our experiments, each row contains images captured by a fixed camera from distinctive perspectives

Multimodal features are extracted by multiple audio-visual signal processing modules in our system. Based on the features, multi-level events are inferred online through our probabilistic reasoning framework. Our reasoning algorithm achieves the speed of 50 fps on Intel Xeon CPU of 2.40 GHz. Output results of some typical state nodes in the EDM-DBN model are presented here based on the test meeting sequence.

State node C_t in the model denotes group situations, and its probabilistic results are given in Fig. 7. Node C_t contains three possible states, "*break*", "*presentation*" and "*discussion*". Probabilities of the three states are inferred simultaneously and the state with the maximum probability is adopted as the output result. As the meeting starts, all the participants start to discuss about meeting arrangement, then a participant goes up and gives presentation. After that, there is a short discussion about the presented lecture, and then participants leave their seats and take breaks. Later participants get back to a long discussion and finally they leave the room.

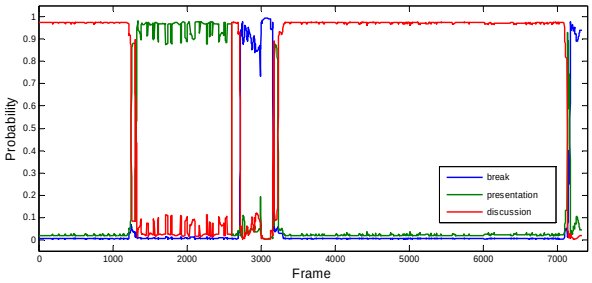

Fig. 7. Output of node C_t

Around frame 3200 there is a false detected "presentation" situation between "*break*" and "*discussion*", which is denoted as the green line in Fig. 8. The three video frames on the right side are extracted at frame 3200, from which we can see that one participant is getting back to his seat. During this period, the other three participants are already standing or sitting at their seats, and the projector is in use. Such kind of "*break*" situation takes the similar characteristics as "*presentation*" scenario, therefore it might incur errors during reasoning process.

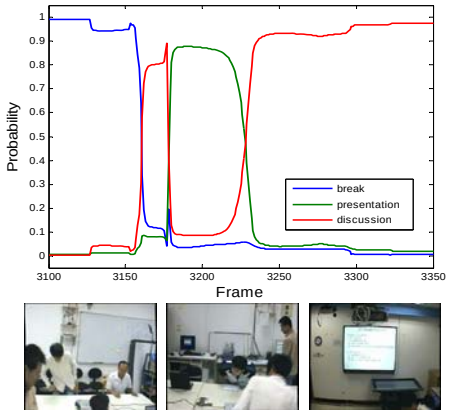

Fig. 8. Segmented result of node C_t's output, video frames from the three cameras correspond to the false detected "*presentation*" scenario in the top graph

Node S_t represents sub-situations and its output results are highly correlated with C_t. Here we analyze the "*asking question*" event detected between frame 2035 and frame 2090, where one of the audiences raises hand and asks a question. It takes place between a "*lecturing*" state and a "*silence*" state, and is depicted with the cyan line in Fig. 9. After the audience finishes the question, the lecturer takes the

turn and answers the question, which is expressed with the red line closely behind the "*asking question*" event. At other intervals, "*lecturing*" states and "*silence*" states switch between each other, which indicate the pauses of the lecturer during "*presentation*".

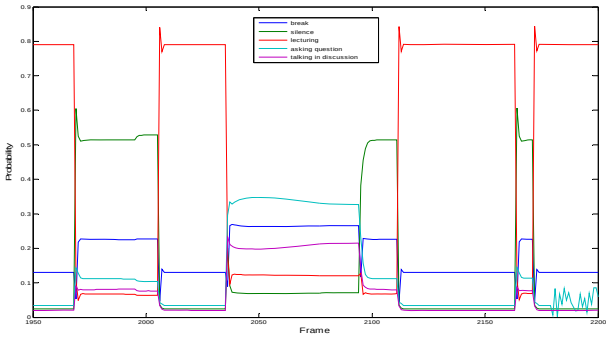

Fig. 9. Segmented result of node S_t 's output. Blue line: "*break*"; green line: "*silence*"; red line: "*lecturing*"; cyan line: "*asking question*"; purple line: "*someone talking in discussion*"

Besides the node C_t and S_t which represent group situations and sub-situations, the node R_t that represents roles is also significant. Here we choose one type of roles R_t^T for the experimental analysis, which represents the speaker locations around the table during "*discussion*". A segment of node R_t^T 's output is shown in Fig. 10, which is extracted during a "*discussion*" scenario. Talker roles during "*discussion*" are assigned to various participants at the four different seats. From the figure we can see a sequence of various interactive patterns. Firstly, participant *D* talks, which is indicated with the cyan line. Later, participants *A* and *C* interact with each other, which are denoted as the blue line and red line. Finally, participants *B* and *C* communicate, which are expressed with the green line and red line respectively in the figure.

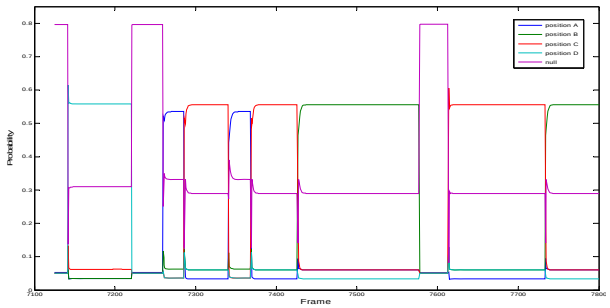

Fig. 10. Segmented result of node R_t^T 's output during "*discussion*". Blue line: location *A*; green line: location *B*; red line: location *C*; cyan line: location *D*; purple line: null

5. 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 of interaction context. This paper presents a novel Event Driven Dynamic Context Model for context-aware computing in meeting scenarios. The essence of context sensing is multi-level event detection, which integrates bottom-up and top-down reasoning mechanism together consistently. A novel probabilistic model EDM-DBN is adopted to model the hierarchical context and events, based on which online detection of multi-level events is achieved. It is characterized as a great advantage of our approach over previous works. The effectiveness of the proposed approach has been tested on the extracted meeting data. Future work includes model improvement, adding more sensors, and integration of proactive services.

Acknowledgments

The work described in this paper is supported by CNSF grant 60673189 and CNSF grant 60433030.

References

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

[2] P. Dourish, and V. Bellotti, "*Awareness and Coordination in Shared Workspaces*", Proc. ACM Conference on Computer Supported Cooperative Work, Toronto, Ontario, ACM Press, 1992.

[3] S. Greenberg, "*Context as a Dynamic Construct*", Journal of Human-Computer Interaction, Vol.16, pp.257-268, 2001.

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

[5] J.L. Crowley, "*Context Driven Observation of Human Activity*", Proc. European Symposium on Ambient Intelligence, 2003.

[6] M.M. Trivedi, K.S. Huang, and I. Mikic, "*Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces*", IEEE Trans. on Systems, Man, and Cybernetics—PART A: Systems and Humans, Vol.35, No.1, pp.145-163, 2005.

[7] J.G. Tan, D. Zhang, X. Wang, and H.S. Cheng, "*Enhancing Semantic Spaces with Event-Driven Context Interpretation*", Proc. PERVASIVE 2005, LNCS 3468, pp.80-97, 2005.

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

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

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