Online Analysis of Hierarchical Events in Meetings

Xiang Zhang[1,2], Guang-You Xu[1], Xiao-Ling Xiao[3], and Lin-Mi Tao[1]

[1] Department of Computer Science, Tsinghua University, 100084 Beijing, China
[2] Yangtze University, 434023 Jinzhou, Hubei ,China
[3] Department of Computer Science and Technology, Wuhan University of Technology, 430063 Wuhan, China
{xiang-zhang,xgy-dcs,linmi}@tsinghua.edu.cn, xljrzx@163.com

Abstract. Automatic online analysis of meetings is very important from three points of view: serving as an important archive of a meeting, understanding human interaction processes, and providing the attentive services based on the meeting situation for participants. Based on this view, this paper presents principle and implementation of online analysis of hierarchical events in meeting scenario. A hierarchical dynamic Bayesian network modeling different levels of events is designed. In this model, the recognition of low-level events is supervised by high-level events Rao-Blackwellized particle filter is proposed for on-line inference for the hierarchical dynamic Bayesian network. Situation events and four sorts of interaction events in meeting scenario are detected and recognized. Experimental results show that our approach can detect and recognize multi-layer semantic events in dynamic environment. Comparing with previous methods of meeting analysis, our approach supports online probabilistic inference for activities at different layers in meeting scenario.

Keywords: Meeting analysis, dynamic Bayesian network, particle filter, event detection and recognition.

1 Introduction

Meetings play an important role in social communication and interaction [1]. Meeting minutes can serve as an important archive of a meeting but they can't provide the attentive services based on the meeting situation for the participants, if the minute is derived off-line after the meeting. Further some important events might be missed. Therefore, it is very important to develop an online meeting archive system capable to analyze the hierarchical events during the meeting, so as to response accordingly. Meeting analysis involving group interaction has attracted attention in fields spanning computer vision, speech processing, human-computer interaction, and information retrieval. Sample applications include structuring, browsing and querying of meeting databases, and facilitation of remote meetings.

Meetings recorded in multi-sensor room consist of multimodal streams of audio and video, captured with multiple cameras and microphone arrays covering participants and workspace areas. The semantic approach is based on representing the meaning of multimodal behavior of a meeting participant using information obtained

N. Aykin (Ed.): Usability and Internationalization, Part I, HCII 2007, LNCS 4559, pp. 472–479, 2007.

from different sources, as well as on recognition of meeting situation actions using semantic features extracted from participants multimodal behavior. Although there are many researches involved in meeting archiving and individual and group actions analysis in meeting rooms, such as Distributed Meeting System from Microsoft [2], Multimodal Meeting Tracker from CMU [3], AMI System from IDIAP [4], and EU research project M4 [5], still few researches tried to incorporate context awareness into meeting event analytical problems. In CMU's CAMEO system [6], personal states and meeting states are inferred through finite state machine models separately, where only users' standing-sitting states are taken into account. McCowan et al. [7, 8] used Layered Hidden Markov Models for the recognition of individual and group actions in meeting scenarios based on audio-visual information. Hakeem and Shah [9] proposed an ontology and taxonomy framework for the offline classification of meeting videos. Al-Hames and Rigoll [10] employed dynamic Bayesian network for the recognition of group actions in meeting video sequences. Most of the research literatures mentioned here performed offline meeting event analysis in predetermined and constrained context models [11,12]. To date most approaches to automatic meeting analysis have been limited to the analysis of the individual actions of meeting participants. Recent work has started to explore multi-person scenarios, where not only individual but also group actions or interactions become relevant.

In this paper, we will describe our semantic approach in modeling a meeting as a sequence of meeting situation events. We propose a hierarchical dynamic Bayesian network for recognition of hierarchical events in meetings. In this model, the recognition of low-level events is supervised by high-level events (regarded as the context). Another novelty of our work here is that we show how Rao-Blackwellized particle filters can be applied to efficiently online estimate joint posteriors over hierarchical events in meeting.

Based on this approach, our paper is organized as follows. Section 2 introduces the meeting room configuration. Section 3 proposes hierarchical events model which we used as hierarchical events structure in the meeting scenario. Section 4 describes the online inference method for hierarchical events. Experimental results are presented and discussed in Section 5, and Section 6 concludes this paper.

2 Meeting Recoding Configuration

The meeting room configuration for events analysis and recordings is illustrated in Fig. 1. Multiple sensors are installed in the meeting room so as to acquire the overall information about the environment in real time. Three fixed cameras are set to extract visual information from three distinct perspectives, where two cameras each acquired a font-on view of two participants including the table region, and the third camera behind the table observes various activities of the participant in font of the meeting room. Three Intel-provided microphone arrays detect the direction of the sound resource and speech activities in real time. The seating positions of four participants were allocated randomly.

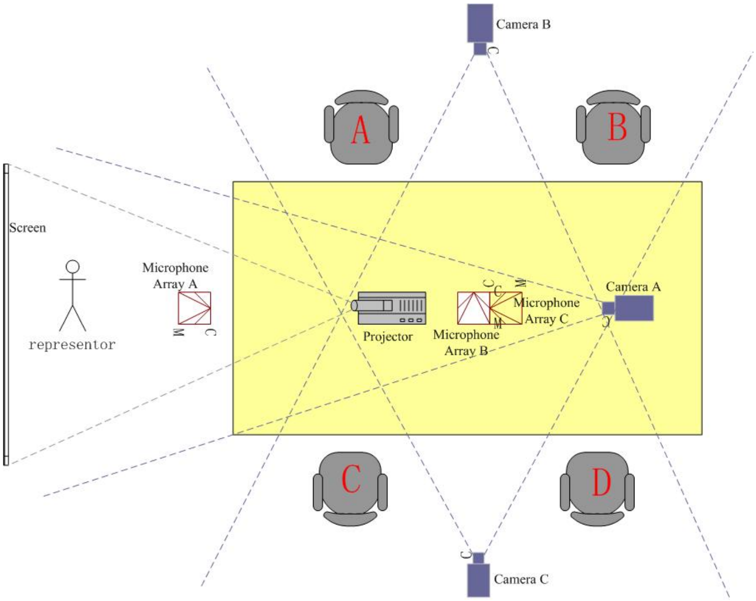


Fig. 1. Meeting recording configuration

3 Hierarchical Events Model

A common drawback in all previously proposed approaches is that they feed the sensor data or features into static classifiers, or a bank of temporally independent HMMs. Further, most of the previously proposed algorithms do not make a distinction between ‘complex’ and ‘simple’ activities [13]. In practice, it might be advantageous to decompose complex activities into simpler activities that might be easier to learn. Our approach divides the meeting actions into three hierarchical components: a set of multimodal group actions (situation events), a set of the interaction actions (interaction events) and a set of individual actions (entity events).

3.1 Definition of Events

A group of four situation events are defined based on multi-modal turn-taking patterns in the meeting scenario. The list is defined in Table 1. These situation events are multimodal, non-overlapping and exhaustive, and commonly found in meetings. There are all natural actions in which participants play and exchange similar, opposite, or complementary roles. For examples, during a monologue, one person speaks to the group, while the other participants listen and direct their gaze towards the speaker or to another one. During a discussion, multiple participants take relatively short turns at speaking, and more movement could be expected.

Four sorts of interaction events during four situation actions are respectively defined, such as motion interaction activities, multi-person speaking interaction activities, appearance of person activities in five room areas, and using the projector screen. Some individual actions consist of sitting, standing, speaking, et al.

Table 1. Description of situation events

Situation events	Description
Monologue	one participant speaks continuously without interruption for a long time
Presentation	one participant at font of room makes a presentation using the projector screen
Discussion	all participants engage in a discussion
Break	each participant is free

3.2 Hierarchical Dynamic Bayesian Network

The dynamic Bayesian network(DBN) allows the construction and development of a variety of models, starting from a simple HMM and extending to more sophisticated models, with richer hidden state [14]. Among the many advantages provided by the adoption of a DBN formalism, one benefit is the unequalled flexibility in the model internal state factorization. Situation events and interaction events in intelligent meeting scenario are detected and recognized here. Two-level hierarchical events, combining with vision and audio feature cues, are modeled using the hierarchical dynamic Bayesian network, as is illustrated in Fig. 2.

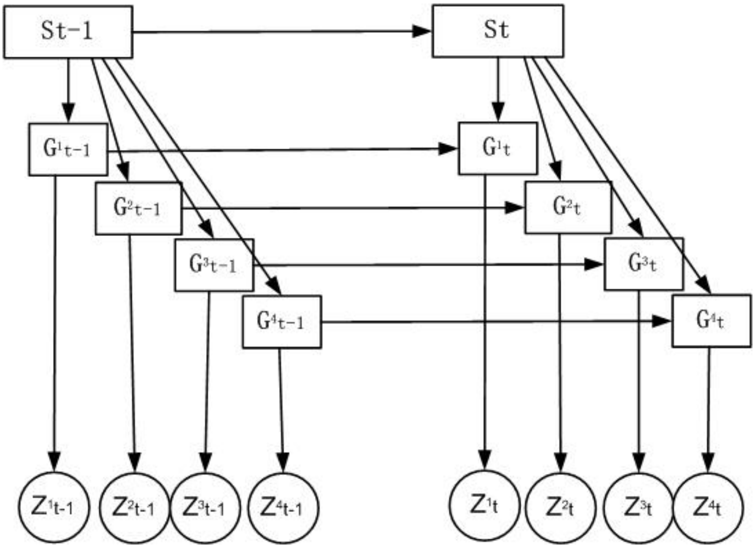


Fig. 2. Dynamic Bayesian network structure for meeting events analysis; square nodes represent discrete hidden variables and circle nodes denote discrete observations

The vision and audio observations of human objects in the five interest areas are treated as observation nodes $Z_t^1 \sim Z_t^4$ of the Dynamic Bayesian Network. Dynamic Bayesian networks estimate four sorts of interaction events $G_t^1 \sim G_t^4$ in the five interest areas from the given observation, and further infer meeting situation events S_t from four cues of interaction events.

4 Inference

Comparing with previous methods of meeting analysis, our approach supports online probabilistic inference for activities at different layers in meeting scenario. During inference, our system estimates a joint posterior distribution over the complete state space of the hierarchical dynamic Bayesian network. Exact solution to this problem will have exponential complexity in the number of levels of the hierarchical dynamic Bayesian network, and thus is intractable.

We describe how Rao-Blackwellized particle filter (RBPF)[15] can be applied for efficient inference in our hierarchical dynamic Bayesian network. Just like regular particle filters, RBPF represents posterior probability over a state space by temporal sets of weighted samples. RBPF derives their efficiency from a factorization of the state space, where posterior probability over one part of the state space are represented by samples, and posterior probability over the remaining parts are estimated exactly, conditioned on each sample.

According to the structure of the hierarchical dynamic Bayesian network in Fig. 2, the joint posterior distribution $p\left(S_{1:t}, G^1_{1:t}, G^2_{1:t}, G^3_{1:t}, G^4_{1:t} \middle| Z_{1:t}\right)$ over the complete state space of the hierarchical dynamic Bayesian network for a sequence of T temporal slices can be decomposed as following:

$$\begin{aligned}
&p\left(S_{1:t}, G^1_{1:t}, G^2_{1:t}, G^3_{1:t}, G^4_{1:t} \middle| Z_{1:t}\right) \\
&= p\left(G^1_{1:t}, G^2_{1:t}, G^3_{1:t}, G^4_{1:t} \middle| Z_{1:t}, S_{1:t}\right) p\left(S_{1:t} \middle| Z_{1:t}\right) \\
&= p\left(G^1_{1:t} \middle| Z_{1:t}, S_{1:t}\right) p\left(G^2_{1:t} \middle| Z_{1:t}, S_{1:t}\right) p\left(G^3_{1:t} \middle| Z_{1:t}, S_{1:t}\right) p\left(G^4_{1:t} \middle| Z_{1:t}, S_{1:t}\right) p\left(S_{1:t} \middle| Z_{1:t}\right) \\
&= p\left(G^1_{1:t} \middle| Z^1_{1:t}, S_{1:t}\right) p\left(G^2_{1:t} \middle| Z^2_{1:t}, S_{1:t}\right) p\left(G^3_{1:t} \middle| Z^3_{1:t}, S_{1:t}\right) p\left(G^4_{1:t} \middle| Z^4_{1:t}, S_{1:t}\right) p\left(S_{1:t} \middle| Z_{1:t}\right)
\end{aligned} \tag{1}$$

Our RBPF algorithm samples the situation event variable S, and computes exact posteriors $p\left(G^1_{1:t} \middle| Z^1_{1:t}, S_{1:t}\right)$, $p\left(G^2_{1:t} \middle| Z^2_{1:t}, S_{1:t}\right)$, $p\left(G^3_{1:t} \middle| Z^3_{1:t}, S_{1:t}\right)$ and $p\left(G^4_{1:t} \middle| Z^4_{1:t}, S_{1:t}\right)$ over the interaction event variables G^1, G^2, G^3, G^4, conditioned on the samples representing situation events.

5 Experiment Results

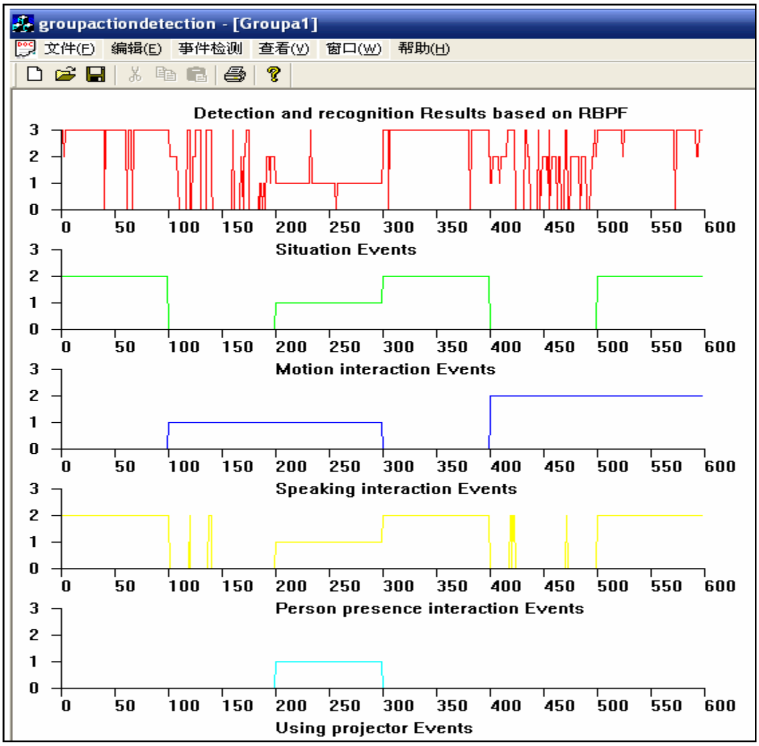
Observation vectors are formed from a range of audio-visual features that measures the individual events. Audio features, the sound source localization, were extracted from the three microphone arrays. Blobs denoting the human body parts, such as the whole body, head, and face, are extracted from video streams, and represented using boxes. We use box attributes such as width, height, and its center position as vision features to estimate the body poses. We make online recognition of situation events and interaction events in meeting using the above hierarchical dynamic Bayesian network.

Model parameters are trained using simple counting methods by the annotated meeting sets. For example, the state transition matrix for situation event S and interaction event G^1, G^2, G^3, G^4, or observation matrix is given by:

$$p(x_j|y_i) = \frac{N(x_j, y_i)}{\sum_{s=1}^{M} N(x_s, y_i)} \tag{2}$$

where $N(x_j, y_i)$ is the number of x_j in y_i, M is the number of state x .

Accuracy was determined by counting the number of correctly labeled frames divided by the total number of frames. Three meeting data sets, which contain four situation events, are annotated. By stochastically selecting two meeting data sets as the training data and the rest as the test data. The total recognition accuracy of situation events and four sorts of interaction events is 85.6%. Fig. 3 shows the detection and recognition results for test data for a sequence of 600 temporal slices.


groupactiondetection - [Groupa1]

文件(F)
编辑(E)
事件检测
查看(V)
窗口(W)
帮助(H)
Detection and recognition Results based on RBPF
Situation Events
Motion interaction Events
Speaking interaction Events
Person presence interaction Events
Using projector Events
0
1
2
3
0
50
100
150
200
250
300
350
400
450
500
550
600

Fig. 3. The detection and recognition results for hierarchical events in the meeting scenario

6 Conclusion

We have described principles and implementation of online analysis of hierarchical events in meeting scenario. We have introduced a hierarchical dynamic Bayesian network that has the ability to model different levels of events and observation features. Rao-Blackwellized particle filter (RBPF) is proposed for on-line inference for the hierarchical dynamic Bayesian network. Some important characteristics of this paper when compared to the previous research works in meeting analysis are: (1) our work can detect not only individual actions but also group or interaction actions, which is very relevant to automatic meeting abstraction; (2) most of the research literatures performed offline meeting event analysis in predetermined and constrained context models, our work can make on-line detection and recognition of human activity in dynamic environment. Experimental results have validated our approach, which show that the RBPF can detect and recognize multi-layer semantic events in dynamic environment.

We are currently exploring several theoretical and engineering challenges with the refinement of hierarchical event analysis in dynamic environment. Context analysis and the relation between context model and event detection are considered for the future extension of our work.

Acknowledgments. This work was funded under Project 60673189 supported by National Science Foundation of China: Event Detection and Understanding in Dynamic Context for Implicit Interaction and Project 2005038351 supported by the Chinese Postdoctoral Science Foundation: On-line Event Detection Based on Context Aware. In addition, we would like to thank the anonymous reviewers for their insightful comments.

References

1. Reiter, S., Rigoll, G.: Multimodal Meeting Analysis by Segmentation and Classification of Meeting based on a Higher Level Semantic Approach. In: Proc. IEEE ICASSP, Philadelphia, USA (2005)
2. Cutler, R., Rui, Y., Gupta, A.: Distributed Meetings: A Meeting Capture and Broadcasting System. In: Proc. ACM Multimedia (2002)
3. Bett, M., Gross, R., Yu, H.: Multimodal Meeting Tracker. In: Proc. RIAO (2000)
4. Carletta, J., Ashby, S., Bourban, S.: The AMI Meeting Corpus: A Pre-announcement. In: Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI) (2005)
5. http://www.m4project.org
6. Rybski, P.E., De la Torre, F., Patil, R., Vallespi, C., Veloso, M., Browning, B.: CAMEO: Camera Assisted Meeting Event Observer. In: Proc. Int. Conf. on Robotics and Automation (ICRA'04), vol. 2, pp. 1634–1639 (2004)
7. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic Analysis of Multimodal Group Actions in Meetings. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI'05), vol. 27(3) (2005)
8. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I.: Modeling Individual and Group Actions in Meetings with Layered HMMs. IEEE Trans. on Multimedia, vol. 8(3) (2006)

9. Hakeem, A., Shah, M.: Ontology and Taxonomy Collaborated Framework for Meeting Classification. In: Proc. 17th Int. Conf. on Pattern Recognition (ICPR'04), vol. 4, pp. 219–222 (2004)
10. Al-Hames, M., Rigoll, G.: A Multi-Modal Graphical Model for Robust Recognition of Group Actions in Meetings from Disturbed Videos. In: Proc. IEEE Int. Conf. on Image Processing (ICIP'05) (2005)
11. Dielmann, A., Renals, S.: Dynamic Bayesian Networks for Meeting Structuring. In: Proc. IEEE ICASSP, Philadelphia, USA (2005)
12. Oliver, N., Garg, A., Horvita, E.: Layered Representations for Learning and Inferring office Activity from Multiple Sensory Channels. Computer Vision and Image Understanding 96, 163–180 (2004)
13. Bui, H.H., Venkatesh, S., West, G.: Policy Recognition in the Abstract Hidden Markov Model. Journal of Artificial Intelligence Research 17, 451–499 (2002)
14. Jensen, F.: An Introduction to Bayesian Networks. Springer, Heidelberg (1996)
15. Murphy, K., Russell, S.: Rao-blackwellised particle filtering for dynamic Bayesian networks. In: Doucet, A., de Freitas, N., Gordon, N.J. (eds.) Sequential Monte Carlo Methods in Practice, Springer-verlag, Heidelberg (2001)