DYNAMIC CONTEXT DRIVEN HUMAN DETECTION AND TRACKING IN MEETING SCENARIOS

Peng Dai, Linmi Tao and Guangyou Xu
Key Lab on Pervasive Computing, Ministry of Education, Tsinghua University, Beijing, China
daip02@mails.tsinghua.edu.cn, linmi@tsinghua.edu.cn, xgy-dcs@tsinghua.edu.cn

Keywords: On-the-spot archiving, dynamic context, human detection and tracking, meeting scenarios.

Abstract: As a significant part of context-aware systems, human-centered visual processing is required to be adaptive and interactive within dynamic context in real-life situation. In this paper a novel bottom-up and top-down integrated approach is proposed to solve the problem of dynamic context driven visual processing in meeting scenarios. A set of visual detection, tracking and verification modules are effectively organized to extract rough-level visual information, based on which a bottom-up context analysis is performed through Bayesian Network. In reverse, results of scene analysis are applied as top-down guidance to control refined level visual processing. The system has been tested under real-life meeting environment that includes three typical scenarios: speech, discussion and meeting break. The experiments show the effectiveness and robustness of our approach within continuously changing meeting scenarios and dynamic context.

1 INTRODUCTION

Context-aware systems incorporate multi-modal information so as to extract semantic understanding of current situation and provide proactive services to users, among which visual information plays a significant role due to its expressiveness and unintrusiveness. Visual detection and tracking of human objects may be a prerequisite for the recognition of human physical and mental states and interactive events, which is essential for those human-centered applications.

Two significant issues lie in the research of context-aware vision systems. Firstly, online analysis of events and context is required so that the context-aware systems might provide services in real-time. Secondly, it is difficult to solve the paradox between low-level visual processing and high-level semantic analysis. In real-life applications context model should be dynamic, which could generate flexible requirements for visual processing mechanism. Therefore context-aware systems must be able to analyze context online, and detect and track human objects in dynamic context. Few research efforts have been attributed to the issue of human detection and tracking in dynamic context.

Meeting room has been demonstrated to be an appropriate research platform for the study of individual and group events analysis (Hakeem and Shah, 2004; McCowan et al., 2005; Zhang et al., 2006). Context model is dynamic during meetings, which not only includes the changes of meeting scenarios, but also covers the changes of individual states or interactive situations. How to deal with low-level feature extraction in such a dynamic context model is a challenging task. Thus we focus our research on online meeting analysis in this paper.

Some researchers have worked on the semantic analysis of meeting video sequences (Hakeem and Shah, 2004; McCowan et al., 2005; Hames and Rigoll, 2005; Zhang et al., 2006), however most of the related work in this domain adopted offline processing frameworks and did not take dynamic context into account, which means high-level context information is not adopted as the online guidance and control of low-level visual processing. An ontology and taxonomy framework was proposed in (Hakeem and Shah, 2004) for the offline classification of meeting videos. Head and hand related events were detected by Finite State Machines based on the tracking result of participants' heads and hands. Those events were further used for meeting scenario classification by rule-based systems. McCowan et al. (McCowan et al., 2005; Zhang et al., 2006) used Layered Hidden Markov Models for the recognition of individual and group actions in meetings based on audio-visual information. Similarly in (Hames and Rigoll, 2005)

Dynamic Bayesian Network was adopted for the recognition of group actions in meeting scenarios. All the approaches mentioned above performed offline context analysis, and visual processing was performed based on predetermined and fixed context.

Recently some research literatures are referring to the research issue of dynamic context in smart environments. A distributed system paradigm was presented for a long-term research of dynamic context capture in indoor environments (Trivedi et al., 2005). Multiple sensors were set in the environment and various visual modules such as human detection, tracking and identification were integrated to extract visual information. However, dynamic context here only denoted changes of user's location and face orientation in the environment, based on which a top-down control was performed for the best perspective selection among multiple cameras. In our work dynamic context is targeted at higher level semantics, including both the individual object events and the overall scenario types.

Previous visual approaches for human presence detection and tracking in meeting environment could not combine various visual cues effectively, hence requirements of real-time and long-term processing in real-life applications could not be matched. Waibel et al. (Waibel et al., 2003) adopted motion and color cues for the segmentation and tracking of human bodies in meeting rooms. Head poses are important cues for the estimation of participants' focus of attention. In (Stiefelhagen et al., 2002), Neural Network models were employed for face detection and pose estimation, Bayesian Network was applied to estimate people's focus of attention. In Hakeem's work (Hakeem and Shah, 2004) tracking algorithms for human head and hands required manual intervention.

In this paper, a novel approach is proposed to solve the problem of context-aware human detection and tracking. A bottom-up and top-down integrated visual processing framework is addressed to solve the paradox between the acquisition of visual cues and dynamic context model. Bayesian Network is adopted to analyze the changing context in a bottom-up way, based on the coarse information of human objects and the meeting room settings. Reversely, refined level information is extracted based on the requirements of current context information.

Additionally, an effective hypothesis-verification approach is proposed to solve the problem of human detection in indoor environment. Potential human objects are detected and tracked based on multiple visual cues, and are further verified with an efficient face detection module.

Thus the major contribution of this paper can be addressed as follows:

- An adaptive visual processing mechanism in dynamic context;
- A hypothesis-verification mechanism for human detection and tracking in indoor environment.

Based on the approach, our paper is organized as follows. Section 2 introduces On-the-Spot Archiving System we used as the research framework. Section 3 describes the hypothesis-verification mechanism for human detection and tracking. In Section 4 a context analysis framework is given, and Section 5 presents the refined level visual modules which can is selectively determined by dynamic context. Experimental results are presented and discussed in Section 5, and Section 6 concludes this paper.

2 ON-THE-SPOT ARCHIVING SYSTEM

Traditional multimedia meeting corpus (Hakeem and Shah, 2004; McCowan et al., 2005) records all the audio-visual information in meetings, which lacks the capabilities of online meeting highlights detection and indexing. In our work, an intelligent multimedia archiving system named On-the-Spot Archiving System (OSAS) is introduced to incorporate multimodal information processing modules into our context-aware research paradigm and has the capability of archiving the most significant information selectively in real-time.

2.1 Research Platform

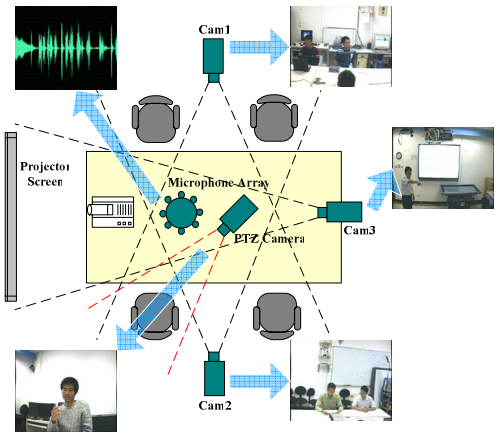

Figure 1: Meeting room settings.

Multiple sensors are installed in the meeting room so as to acquire the overall information about the environment, as is illustrated in Figure 1. Three fixed cameras are set to extract video frames from distinct perspectives. In this paper, visual processing and context reasoning is constrained on the video sequences acquired with the 3 fixed cameras.

In our research experiments, the problem is simplified by restricting the meeting participant number to be 4. Three typical sorts of meeting scenarios are taken into account in this paper: speech, discussion, and break, as described in Table 1 below.

Table 1: Three meeting scenarios.

Scenario	Features
Speech	One participant gives speech at the projector screen, the other three seated.
Discussion	Four participants seated and talk.
Break	Participants leave the seats and perform random actions.

On-the-Spot Archiving system is designed to analyze individual and group events in the meeting room, and archive the multimodal information appropriately based on the online analysis results of meetings. Archived multimodal information can be retrieved and browsed by users later. The prototype of our system is illustrated in Figure 2.

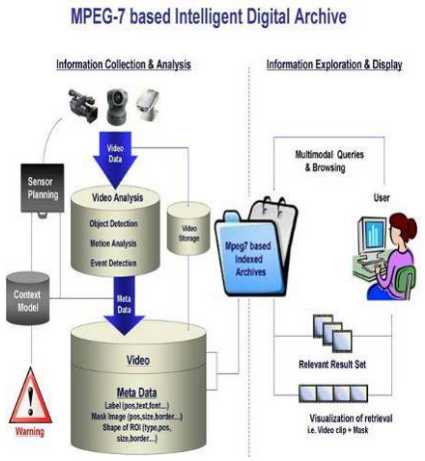

Figure 2: Prototype of on-the-spot archiving system.

Visual processing in a human-centered smart environment can be classified into two major categories: (1) coarse level visual modules, i.e. detection and tracking of head and body blobs; (2) refined level visual modules, including detection and tracking of hand blobs, head pose estimation etc.

The objective of our system is to record the most relevant information according to dynamic context. For instance, before the meeting begins or during meeting breaks, only the coarse level information processing modules are required. While during speech or group discussion, refined level visual processing is required so that we can analyze the participants' actions and focus of attention.

2.2 Context-Aware Visual Processing

All participants need to be analyzed via visual processing techniques so as to recognize individual and group events. While at the same time, analysis of individual activities and group events in different meeting scenarios might demand distinct visual cues.

Thus a novel framework is proposed to solve this issue, which proceeds the context-aware visual mechanism in a loop way. Firstly, coarse level human blobs are detected and verified so as to understand current meeting scenarios in a bottom-up mode; then in a top-down fashion, current meeting context model helps control the selective processing of refined level blob information, as is illustrated in Figure 3.

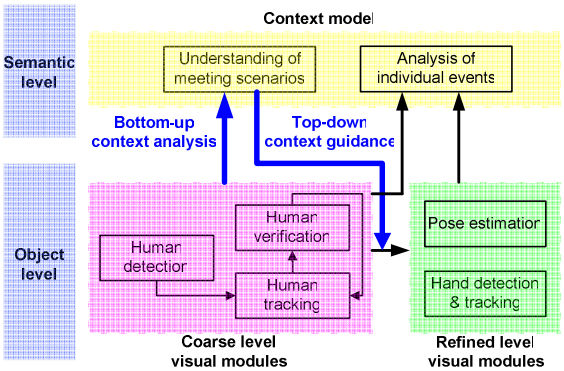

Figure 3: Context-aware vision system flowchart.

3 HUMAN DETECTION AND TRACKING

Human detection and tracking generates coarse information for the bottom-up context analysis, which is a significant starting point for context-aware vision systems.

In this paper a two-stage approach is presented to detect and track potential human objects within meeting scenarios. The entire procedure can be divided into two stages: hypothesis generation and human object verification, which are to be described in details respectively below.

3.1 Hypothesis Generation

Multiple cues are combined to generate hypothesis for potential human objects in the meeting room environment. Motion information is used firstly for foreground object extraction. Color and gradient information is further employed to determine potential head objects. Once the head candidates determined, body blobs are generated according to somatological knowledge.

Motion detection or foreground extraction is the first step of our human detection approach, since motion is a significant clue of human presence, and by foreground extraction we can immensely reduce the search range for further human detection modules. In this paper, a behavior analysis oriented consistent foreground detection method (Jiang et al., 2006) is employed to generate moving objects in indoor environment. Unlike traditional Gaussian mixture model (GMM) based methods, this algorithm maintains two background models, the original background and the run-time background, based on which foreground and background pixels are separated effectively. Figure 4 shows the result of our method compared to traditional GMM results.

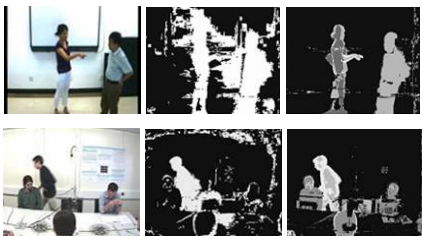

Figure 4: Comparison of GMM based foreground detection and the adopted method here.

Connected component analysis is performed on the detected foreground pixels, which generates a set of foreground objects. Based on these detected objects, color and gradient cues are further processed so as to determine potential head location.

Skin detection is performed in the upper half part of the foreground objects. Skin color model λ is based on HSV color space, and skin and non-skin pixels are classified according to the distribution of hue and saturation, as is expressed in Formula 1.

$$p(I_{(x,y)} \mid \lambda) = p(h_{(x,y)} \mid \lambda)\, p(s_{(x,y)} \mid \lambda) \qquad (1)$$

As a complementary cue of skin color information, object gradient information is also used to determine exact boundaries of human heads. In a neighbor region of skin areas, a set of ellipses $X_i = [x_i, y_i, w_i, h_i]$ $(i = 1, \ldots, N)$ are generated based on changing locations (x_i, y_i) and changing sizes (w_i, h_i) within a certain range. Potential head candidate is determined according to the observation probabilities:

$$X = \arg\max_{X_i} p(y^{edge} \mid X_i) \qquad (2)$$

Figure 5 shows the result of head candidates generation by skin detection and elliptical fitting. Body blobs are further determined based on motion detection results and somatological knowledge concerning the relationship between head and body.

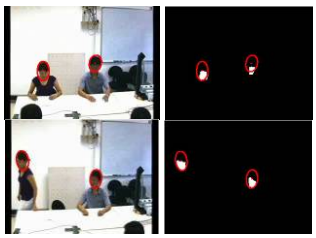

Figure 5: Results of head candidates generation.

3.2 Tracking and Verification

Detected candidate head ellipsis and body rectangles are tracked by particle filtering techniques. Elliptical head state vector is defined as $X_f = [x_f, y_f, w_f, h_f]$, where (x_f, y_f), w_f and h_f denote the center and two axes of the head ellipse respectively. Rectangular body state vector is defined as $X_b = [x_b, y_b, w_b, h_b]$, where (x_b, y_b), w_b and h_b denote the center, width and height of the body rectangle respectively. Either for head or body tracking, a set of samples $\{x_t^{(i)}, \pi_t^{(i)}\}_{i=1}^{N}$ are maintained and updated, and the mean state of the sample set can be adopted as the estimated state of the current object.

Traditional particle filter methods cannot deal with long-term tracking applications effectively due to their lack of appropriate adjustment mechanism. Therefore we propose a revised particle filter tracking algorithm so as to reduce the potential tracking failures. Once the tracking samples drift away from the object for a certain threshold η, head and body detection results are adopted for a self adjustment of the sample set. Those detected head or body blobs overlapping with currently maintained samples are used for the re-initialization of the sample set. Details of our self-adjusted particle filter tracking algorithm is given in Table 2 below.

Through the previous tracking approach, a set of hypothesized human objects are generated and updated, based on which an efficient AdaBoost face detector (Wang et al., 2006) is adopted to verify those objects in video streams.

Table 2: Self-adjusted PF tracking algorithm.

With the particle set $\{x_{t-1}^{(i)}, \pi_{t-1}^{(i)}\}_{i=1}^{N}$ at the previous time step, proceed as follows at time t:

Step 1: **Resampling**. Resample $\{x_{t-1}^{(i)}, \pi_{t-1}^{(i)}\}$ to get $\{x'^{(i)}_{t-1}, \pi'^{(i)}_{t-1}\}$ based on probability $\pi_{t-1}^{(i)}$

Step 2: **Prediction**. Propagate each sample $x_{t-1}^{(i)}$ by a linear dynamic model $x_t^{(i)} = Ax'^{(i)}_{t-1} + w_{t-1}^{(i)}$, where $w_{t-1}^{(i)}$ is a Gaussian random variable

Step 3: **Observation**. For each sample $x_t^{(i)}$, calculate its weight according to color and edge cues:
$\pi_t^{(i)} = p(y_t^{color} \mid x_t^{(i)}) p(y_t^{edge} \mid x_t^{(i)})$

Step 4: **Estimation**. Estimate the mean state of the set $\{x_t^{(i)}, \pi_t^{(i)}\}_{i=1}^{N}$ by $E[x_t^{(i)}] = \sum_{i=1}^{N} \pi_t^{(i)} x_t^{(i)}$

Step 5: **Adjustment**. If the highest weight of the sample set $\pi_{t\,\max}^{(i)} < \eta$, adjust head or body object based on detection results and restart the sampling procedure

AdaBoost face detectors integrate a set of weak classifiers and group them in a pyramid structure to generate a strong classifier. The selection of features and weights is learned through training procedure.

$$H(x) = \begin{cases} 1, & if \sum_{i=1}^{N} \alpha_i h_i(x) \geq (\sum_{i=1}^{N} \alpha_i)/2 \\ 0, & otherwise \end{cases} \quad (3)$$

Besides traditional symmetric rectangle features, asymmetric rectangle features are also adopted in the AdaBoost learning algorithm so as to detect multi-view faces, as is illustrated in the left part of Figure 6. Asymmetric rectangle features can interpret asymmetric gray distribution in profile face image.

In this paper, an appropriate sub-window is generated within the neighbour area of tracked head ellipse. AdaBoost face detector is employed to search face rectangles in this sub-window. Once the face detected, the target object is certified to be valid. As a result of the object layer, input image sequences are converted into sequences of human objects, which are adopted for the following context analysis tasks.

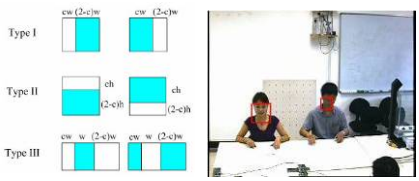

Figure 6: Multi-view AdaBoost face detector (Left: Rectangle feature set of our approach; Right: Face detection result).

4 ANALYSIS OF DYNAMIC CONTEXT

Context analysis is performed on-line in our On-the-Spot Archiving System and applied as the guidance of refined visual modules, which is the most prominent feature between our framework and most previous related work.

In this paper, the concept of dynamic context lies in the changing of meeting scenarios. Thus context analysis here is confined to the recognition of meeting scenarios based on the extracted visual cues. According to the meeting room settings introduced previously, five interest areas are defined in the three camera views, as is illustrated in Figure 7. Human presence and their standing-sitting states in these five areas are adopted as inferring cues.

Figure 7: Interest areas in meeting environment.

Bayesian Network is implemented to estimate human presence status in the five interest areas and meeting scenario at each time step. The observations of human objects in the five interest areas are treated as observation nodes $O_1, \cdots, O_5$ of the Bayesian Network. Bayesian Network estimates human presence states $S_1, \cdots, S_5$ in the five interest areas from the given observation and further infer meeting scenario S_m from these individual states:

$$S_m = \arg\max_j P(S_m = j, S_1, \cdots, S_5 \mid O_1, \cdots, O_5) \quad (4)$$

Those individual and group states in BN are defined in discrete values as follows:

$$S_i = \{\text{nobody, standing, sitting}\} \ (i = 1, \ldots, 5) \quad (5)$$

$$S_m = \{\text{speech, discussion, break}\}$$

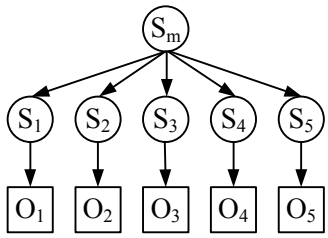

Figure 8: Bayesian Network structure for meeting scenario understanding.

Figure 8 illustrates the basic structure of the Bayesian network. S_m is the final output of the context analysis, which classified current meeting

context into three scenarios: speech, discussion and meeting break. As a result of the context analysis layer, meeting video sequences are converted into a sequential outline of meeting scenarios.

Current work contains no temporal information within our reasoning framework, Dynamic Bayesian Network is considered to be adopted in future work.

5 CONTEXT-DRIVEN REFINED VISUAL PROCESSING

In some specific meeting scenarios such as speech and discussion, detection, tracking and analysis of detailed blobs are required so as to provide more refined cues for further analysis of individual events.

5.1 Context-Aware Visual Processing

As we can see, video sequences before, in-between or after a complete meeting are all categorized into meeting break scenarios. During meeting breaks, no significant information is generated or propagated about the meeting contents, therefore there is no need for refined level visual processing.

On the contrary, those meaningful individual and interactive events generated during speech or discussion scenarios contain the majority of meeting information, hence refined level visual modules such as head pose estimation and hand tracking are required for the further analysis of human activities and interactions.

As is described in Table 3, head pose estimation is applied to help analyze participants' focus of attention and their concentration level toward the meeting. Hand detection and tracking module can be used for the recognition of specific human actions such as raising hands and pointing directions.

Table 3: Detailed-level visual processing for specific meeting scenarios.

Visual module	Scenario	Objective
Pose estimation	Speech, discussion	Analyze focus of attention.
Hand tracking	Speech	Analyze hand actions such as raising hand, pointing and speech related actions.

5.2 Head Pose Estimation

A Bayesian estimation method (Park, 2004) can be further employed to recognize participants' head poses rough.

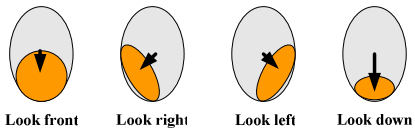

Figure 9: Head pose estimation.

Denote the state of head pose in the current image with X . Two types of rough observations are selected as reasoning cues, let Y_a be the angle of vector from head center to face center, Y_r be the ratio of face area to head area, as is described in Figure 9. Pose estimation task is equivalent to estimate the maximum belief:

$$P(X \mid Y_a, Y_r) \propto P(Y_a \mid X) P(Y_r \mid X) \qquad (6)$$

5.3 Hand Detection and Tracking

Use the same skin color detection technique and connected component analysis algorithm as in Section 3, hand blobs can be segmented successfully.

Figure 10: Hand detection and tracking.

Tracking of hand blobs appeals to color based Mean-Shift algorithm. We apply skin color model introduced in Section 3 to the human body areas and generate probability maps, based on which a Mean-Shift algorithm is carried out to track hand rectangles. Selected samples of hand detection and tracking in our system are presented in Figure 10.

The output of hand detection and tracking is a trajectory of hand blobs, which can be used as input for the recognition of more individual events, such as raising hand for QA and object-targeted pointing.

6 EXPERIMENTS

We apply the proposed approach to detect and track human in our meeting room environment. Figure 11 shows the visual processing results performed with each of the three camera views during a meeting.

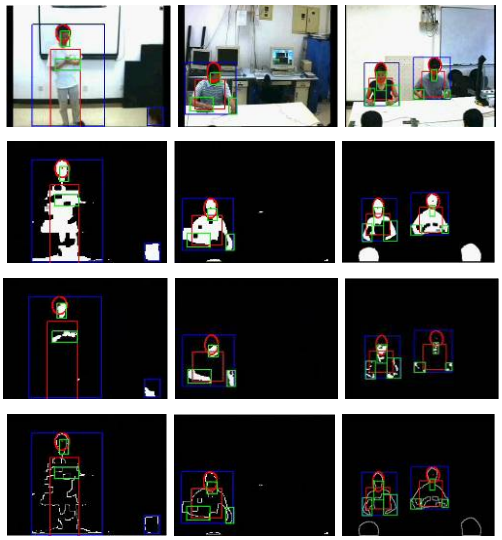

Figure 11: Visual processing and human object detection results (1st row: original images; 2nd row: foreground detection; 3rd row: skin detection; 4th row: Edge detection).

Our approach has been demonstrated with real data extracted in our meeting room. Figure 12 shows the result of context analysis from one of the meeting samples. Meeting scenario recognition rate is 95.4%. Most of the false alarms come from those false detected 'speech' scenarios. For instance, three pink marked sections expressed in Figure 12 are typical false detected samples. Such types of errors are generated due to insufficiency of our context analysis model. In our current model, such meeting scenes with one person standing or walking at the projector screen and the other three seated around the table are regarded as 'speech' scenarios. However, sometimes such kind of scenes might also happen during meeting breaks. Recognition rate can be improved by incorporating more cues and modifying context analysis model in the future.

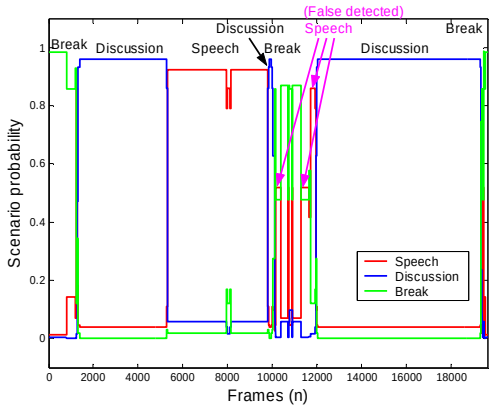

Figure 12: Context analysis result.

Figure 13 gives human tracking results before the meeting starts. As we can see, only coarse level head and body blobs are tracked during this period. Initially only one person is detected since two persons are detected as one foreground object. Once the two separated and are detected as foreground objects respectively, the other one is detected.

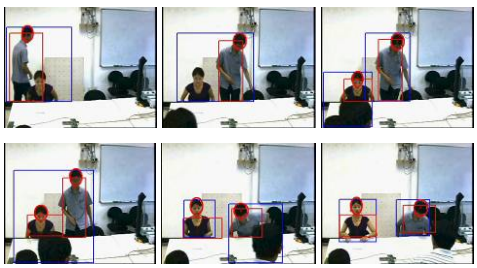

Figure 13: Tracking results before starting of the meeting, in which head pose estimation and hand tracking are not performed.

Experimental results shown in Figure 14 demonstrate the effectiveness of our approach in human verification while the presenter walks back to his seat. The initially false detected head blob is eliminated through verification procedure.

Figure 14: Human verification and tracking results while the presenter is walking back to his seat and sitting down.

Figure 15 and 16 indicate the effectiveness of our context-aware visual processing approach. During the speech scenarios, both the coarse level blobs and the refined level blobs are tracked. Figure 15 shows the tracking results of the participant's raising hand process, and the speaker is tracked stably for a long time span in Figure 16.

Figure 15: Tracking results while the participant is raising his right hand.

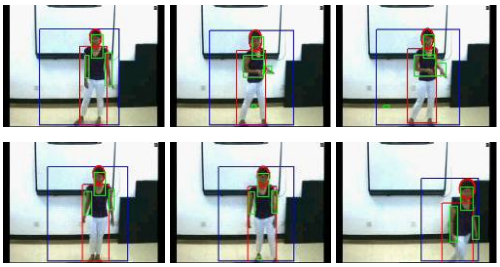

Figure 16: Tracking results of the presenter.

However, our tracking algorithm might generate errors when two participants walk across each other and one of the heads is occluded for a short period. Cooperative reasoning structure is considered to improve our approach in future work.

7 CONCLUSIONS

In this paper, a bottom-up and top-down integrated visual framework is proposed for human-centered processing in dynamic context. Coarse level visual cues are extracted concerning human presence and states in meeting scenarios, based on which context analysis is performed through Bayesian reasoning approach. Context information is then applied to control refined visual modules in a top-down style. Besides, a novel hypothesis-verification method is adopted for robust detection and long-term stable tracking of human objects. Experimental results have validated our approach. Spatial-temporal analysis of hierarchical context model is considered for the future extension of our work.

ACKNOWLEDGEMENTS

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

REFERENCES

McCowan, I., Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D., 2005. Automatic Analysis of Multimodal Group Actions in Meetings. In *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI'05)*, Vol. 27, No. 3.

Zhang, D., Perez, D., Bengio, S., McCowan, I., 2006. Modeling individual and group actions in meetings with layered HMMs. In *IEEE Trans. on Multimedia*, Vol. 8, No. 3.

Waibel, A., Schultz, T., Bett, M., Denecke, Malkin, R., Rogina, I., Stiefelhagen, R., Yang, J., 2003. SMaRT : The Smart Meeting Room Task at ISL. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing 2003 (ICASSP'03)*, Vol. 4: 752–755.

Stiefelhagen, R., Yang, J., Waibel, A., 2002. Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues. In *IEEE Trans. on Neural Networks*, Vol. 13, No. 4.

Trivedi, M., Huang, K., Mikic, I., 2005. Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces. In *IEEE Trans. on Systems, Man, and Cybernetics—PART A: Systems and Humans*, Vol. 35, No. 1.

Hakeem, A., Shah, M., 2004. Ontology and taxonomy collaborated framework for meeting classification. In *Proc. 17th Intl. Conf. on Pattern Recognition 2004 (ICPR'04)*, Vol. 4: 219-222.

Hames, M., Rigoll, G., 2005. A Multi-Modal Graphical Model for Robust Recognition of Group Actions in Meetings from Disturbed Videos. In *Proc. IEEE Intl. Conf. on Image Processing 2005 (ICIP'05)*.

Song, X., Nevatia, R., 2004. Combined face-body tracking in indoor environment. In *Proc. 17th Intl. Conf. on Pattern Recognition 2004 (ICPR'04)*, Vol. 4: 159-162.

Wang, Y., Liu, Y., Tao, L., Xu, G., 2006. Real-Time Multi-View Face Detection and Pose Estimation in Video Stream. In *Proc. 18th Intl. Conf. on Pattern Recognition (ICPR'06)*, Vol. 4: 354-357.

Jiang, L., Zhang, X., Tao, L., Xu, G., 2006. Behavior Analysis Oriented Consistent Foreground Object Detection. In *Proc. 2rd Chinese Conf. on Harmonious Human Machine Environment 2006 (HHME'06)*.

Park, S., 2004. A Hierarchical Graphical Model for Recognizing Human Actions and Interactions in Video. *PhD Thesis*.