Active video recording system for meeting

Ing-Jr Ding*, Chi-Yuan Wang **, and Kun Cheng Tsai+
*, **, + *Networks and Multimedia Institute, Institute for Information Industry, Taipei, Taiwan*
* ingjr@nmi.iii.org.tw, ** richwang@nmi.iii.org.tw, +garytsai@nmi.iii.org.tw

Abstract

An Application is developed on video and audio information based environment for an enterprise, holding a meeting. In addition, we demonstrate real-time enterprise that facilitates effective and efficient meetings with the help of an active meeting room. We have developed techniques for location-aware and behavior modeling using sensor data. This system can adjust meeting equipments based on the interactions, among the participants in the room and automatically create a searchable audiovisual record summary of the meeting including audio and video clips. Finally, an audiovisual review program can allow the participants to review and the others to share the discussions during and after the meeting.

1. Introduction

It has been recognized that communications between small embedded sensors and related data processing devices are integral facilities in many ubiquitous computing scenarios. It is a common knowledge that context-aware services [1] involves a huge amount of spatial and other contextual information to be sensed, exchanged and processed along with other information from a number of other pervasive devices. Such ubiquitous computing and communication have opened many great opportunities to provide novel solutions to various issues in real human life.

Meeting is one common human activity and consumes a lot of time/energy. During meeting many activities can be seen, tremendous data might be produced if we keep track of all attendants without selection. Thus, we developed a system, to determine whether or not, to take action for the devices in a meeting based on person interaction. Also, we can get the position of each attendant by using microphone arrays through Sound Source Localization (SSL) system.

2. Locating and Recognizing Sound Events in the Meeting

Since the early 1990's, many related works have been focused on how to acquire, refine, and use location context information. In our works, the localization system takes use of sounds to track members in a meeting. When acquiring the location of sounds, this will be helpful for some kind of context aware computing, such as speaker identification. The following scenario describes our designed location-sensing system which locates and recognizes sound events in a meeting using microphone array techniques.

2.1. Sound Source Localization

An SSL system determines the location of sound sources based on the audio signals received by an array of microphones at different known positions in the environment. All microphones receive time-shifted signals mixed with environmental noise and reverberation. Based on current SSL research literature [2], the main challenges for deploying SSL systems in domestic environment are: 1. Background noise 2. Reverberation (echoes) 3.Broadband 4.Intermittency and Movement 5. Multiple Simultaneous Sounds.

Despite these general challenges mentioned above, our research system shows that it is feasible and useful to start investigating how sound source location can be identified, which drive a camera to focus on locations where sound events are detected.

Different effective algorithms with an array of microphones are used in sound source localization. They can be divided into some main categories [2]. Most current SSL systems are based on computing Time-of-Delay (TOD), and our SSL system follows that. The detail descriptions are as follows:

Consider this case, a person speaks normally, therefore audio features using multiple microphones to detect speaker position can be utilized. Consider the j-th pair of microphones, and let m_{j1} and m_{j2} respectively be the positions of the first and second microphones in the pair. Let x denote the position of the speaker in a three dimensional space. Then the time delay of arrival (TDOA) between the two microphones of the pair can be expressed as

$$Tj(x) = T(m_{j1}, m_{j2}, x) = \frac{\| x - m_{j1} \| - \| x - m_{j2} \|}{c} \quad (1)$$

Here " c " represents the speed of sound.

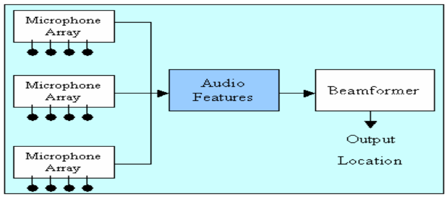


Figure 1. The architecture of the SSL system

To estimate the TDOAs, a variety of well-known techniques [3, 4] can be used to solve it. In our work, a simple delay and sum beamformer is used to locate the sound event. Figure 1 shows the architecture of the SSL system.

2.2. Audio feature selection

Feature extraction is made on 20 ms analysis frames with a 50% overlap. Computed features are chosen among the most popular in audio processing algorithms and more likely to fit with our location problem. Short time energy describes signal energy at a given time and is alternatively referred to as loudness or volume. The first eight Mel-Frequency Cepstral Coefficients (MFCCs) are used. The first two spectral statistical moments, namely the spectral centroid which is the mean of power spectrum for a given time and the spectral spread are used. The first and second derivatives of each of the above features are then used as audio features. In order to reduce the size of the audio feature vector, its dimension is then reduced by principal component analysis procedure. Finally, we keep the first 13 components as significant. Each analysis frame of the input audio signal is thus represented by a 13-dimensional vector.

2.3. An important application of a smart meeting: speaker identification

An important aspect of smart meeting applications is speaker identification. Several additional possible biometric features can be considered for a smart meeting application, but most of them are impractical due to privacy issue. To classify the member in a meeting according to the acoustic speech features is one more feasible option.

Speaker identification is implemented by analyzing short-time spectrum of spoken phrases. In speaker identification, the Gaussian Mixture Model (GMM) has been proven to be a good choice to capture speaker's information in (MFCC). Our designed system hires this kind of model to do works of the speaker's classification.

3. Active Meeting Information System

We aim to demonstrate all enterprises which can facilitate effective and efficient meetings with the help of an active meeting room. An active meeting room refers to intelligent space equipped with active (pan) cameras, microphone arrays, video projector, projector screen, whiteboards, and wireless networks. An example is shown in figure 2 for setting of active meeting room. These equipments are connected and managed by a central active meeting information system (AMIS).

For example, AMIS automatically adjusts the direction of the camera toward the people who is speaking in the room.

By recognizing the activities, the system can track the participants in the meeting room using the data provided by sensors, analyzes sensor data and determines the activity in the room, evaluate the given rules to see if certain actions should be triggered to change the room setting, generate indexed searchable meeting records and store in an audio/video database.

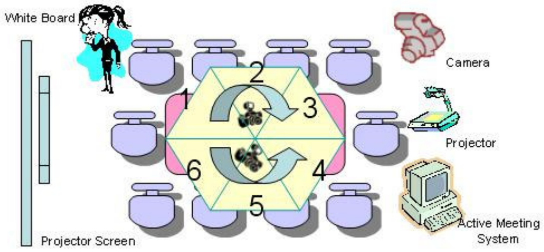


Figure 2. Active Meeting Information System

4. Activity Recognition

AMIS system consists of the activity recognition module. This module recognizes activities primarily based on meeting activity model and the events detected in the meeting room environment. Sec 4.1 describes how Meeting activity model helps in activity recognition and Sec 4.2 describes how event is recognized.

4.1. Meeting activity model

Meeting event ontology plays an important role in activity recognition. Ontology can provide information, such as possible location, time and frequency of event, corresponding to various different events. The system can recognize activity more precisely with its help. Sequential activity model does not recognize activity as separate task. On the other hand, it sees activity as a sequence of flow. Thus the activity done in earlier hour can impact the possibility of activity done later.

4.2. Event-driven recognition

The system recognize activity consists of three event-detection module, Speaking events detection module, Topic events detection module and Movement events detection module, and Active region tracking module.

Speaking events was detected by using the mote wear on the attendant. The motes are for processing and transmitting the audio data. It is likely to provide more functionality such as speech to text or speaking recognition. Different activities usually result in different speaking events, in different level. For example, most speaking events come from one person when the speaker is presenting his slide. On the other hand, more people are involved in if it is time for discussion.

Topic event detection module is done through one power-point plug-in. The plug-in collects information about the slide, and provides useful information for providing context-aware meeting record. The activity recognition module takes advantage of Movement events detection module as well. With the information about movement event such as event trigger time, event trigger place and frequency of movement event, we can determine the active level of meeting. For example, movement event happens mostly at door when the meeting is just going to start and happens at one region when the people are crowded and discuss.

5. System Architecture

The meeting room is equipped with six cameras, each of which is paired with one embedded computing device. In the meeting room, there is one projector screen, one remote-controlled projector, and one computer as server.

To capture video of attendants who may appear in different place of the meeting room, are based on three assumptions;

(1) Each attendant should wear one mote with acoustic sensor and RF radio that can collect the surrounding context data from the aforementioned pervasive devices. Sometimes, the amount of data gathered by the attendant's devices is enormous.

(2) We assume that each mote will appear in one pre-defined region in the room.

(3) Instead of using complex activity models, the system considers only active regions, which do not involve temporal relation between events.

In designing AMIS system, these fundamental requirements are essential to build rational and feasible system architecture. To smoothly perform efficient operations, data must be semantically represented with some metadata [5] and the amount of data has to be greatly reduced by abstracting only the most needed information, to describe meaningful speaking events.

6. System Requirements

Hardware and software requirements are described as follows:

6.1. Hardware requirement

The hardware requirements include devices to control the facility, to track attendants, and to collect activity information from the meeting room. The devices are described as follows:

6.1.1. Directional Microphone

Audio is part of the meeting record in this project. To record audio during meeting, we use directional microphones for recording. Directional microphones can reduce noise from undesired direction, and perform high quality audio record. One directional microphone is used for each pre-defined region. When one region is identified as an active region, corresponding directional microphone will be activated to record audio in the active region.

6.1.2. Motes

To identify speaking participant, system uses a Berkeley Motes sensor kit (figure 3) as a necklace hung in front of a participant's chest to detect the sound volume. Berkeley Motes sensor kit contains a microphone to transfer sensed sound volume to digital values. Louder the sound sensed by the microphone, larger vibration of values will be recorded. The system uses recorded values to estimate speaking events.

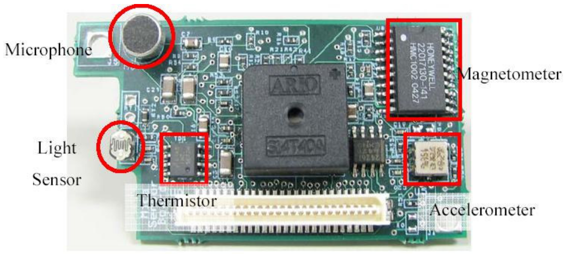


Figure 3. Multi-Sensor Board MTS310

6.1.3. Video Analysis Server

Video Analysis Server (VAS): figure 4 illustrates the VAS which has Windows CE.NET4.1 installed on it.

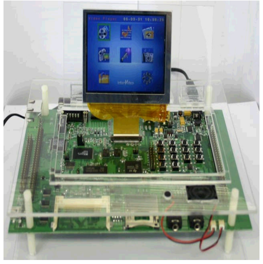


Figure 4. Video Analysis Server Figure 5. Camera

6.1.4. Logic Tech QuickCam Sphere

Logic Tech QuickCam Sphere is used for the panning camera (figure 5).

6.2. Software requirement

The prototype is developed under Microsoft Windows platform. The development tools used in this project are as follows: Microsoft Embedded Visual C++ 4.0, C#, C++, TinyOS 1.1.0 IS, and MySQL. In addition, other SDKs are used in the development: Intel OpenCV, Microsoft DirectX SDK 9.0c, and Microsoft .NET Framework SDK Version 1.1, JDK 1.4, and JavaCOMM 2.0.

6.2.1. Facility control

We use X10 and IR module to control room facilities. The X10 can receive signal from electric circuit and adjust its output voltage. Therefore, if the illumination of meeting room lights can be changed with voltage, we can embed X10 into the power supply circuit of meeting room lights to dim and to lighten the light. After recording the IR message of projector remoter into IR module, we can control the projector with the IR module.

6.2.2 Object tracking and image capturing

The object tracking is used in the system to detect events and capture video streams. In particular, the module is designed to detect if an attendant moves in or out a region. Our approach is based on three-layer view, as illustrated in figure 6. The task layer is responsible for observing the attendant behavior. When an event is triggered, the task layer forwards event to module layer. Then the module layer handles the subscription of video record. When image capturing module receives a subscription request from server, the module will capture video streams and periodically send streams to server.

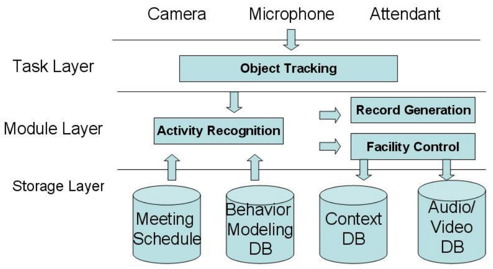
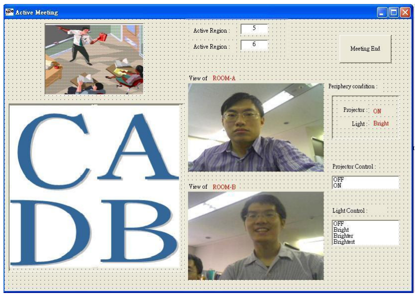


Figure 6. three-layer views

7 Meeting Recording and User Interface

Meeting Recording Module is responsible for recording video and audio of the meeting room instantly. User can run this application in a PC. As shown in figure 7, The GUI displays brightness control and project control which allow changes. As soon as user clicks the option "Meeting begins", meeting progress is recorded. The user can shift the view to any of the door areas "A, B, C, D" by clicking blue text "A, B, C, D". After activation images are obtained from Room A and B.



Active Meeting

Active Region : 5

Active Region : 6

Meeting End

View of ROOM-A

Periphery condition :

Projector : ON

Light : Bright

Projector Control :

OFF
ON

View of ROOM-B

Light Control :

OFF
Bright
Brighter
Brightest

Figure 7. User Interface of Recording Program

8. Review Module

After completion of meeting, when user wants to review the meeting, using Review Module we can review both audio and video which is stored in the data base, and also can be reviewed whenever required. With CSVM edit mode, we can edit audio clips, cut off unnecessary clips or merge several clips together into one. As illustrated in figure 8.

Figure 8. CSVM (edit) Figure 9. CSVM (play)

CSVM play mode can also play the slides and show it to another panel with multiple types of file, such as html, ASF, AVI, WMV, and Flash. As illustrated in figure 9.

9. Conclusion and Future Work

We design activity recognition algorithm to identify certain pre-defined activities, demonstrate our location-aware algorithms and build the application with graphical user interface. For the future enhancement work, we like to, firstly, provide virtual clock synchronization over sensor data streams. Since many sensors are involved in an active meeting room system. To recognize correctly the events or activities, sensor data streams have to be synchronized in terms of time. Secondly, conduct data fusion over sensor nodes. In current design, most of the sensor data are processed on server nodes. However, if the number of sensor nodes increase, such design might suffer the performance issue. We need to aggregate the data on intermediate nodes so that workload on servers can be reduced.

12. References

[1] Martin Bauer, Christian Becker and Kurt Rothemel, "Location Models from the Perspective of Context Aware Applications and Mobile Ad Hoc Networks," Personal and Ubiquitous Computing, Vol. 6, No. 5,
December 2002.

[2] DiBiase, J., Silverman,H., and Brandstein, M., "Robust Localization in Rever-berant Rooms, in Microphone Arrays: Signal Processing Techniques and Applications," M.S.B.a.D. B.Ward, Editor. 2001, springer.

[3] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," Proc. ICASSP, vol. II, pp. 273-6, 1994.

[4] J. Chen, J. Benesty, and Y. A. Huang, "Robust time delay estimation exploit-ing redundancy among multiple microphones," IEEE Trans. Speech Audio Proc,vol. 11, no. 6, pp. 549-57, November 2003.

[5] Catriel Beeri, Tova Milo, "Schemas for Integration and Translation of Structured and Semi-Structured Data," International Conference on Database Theory, 1999.