IEEE Workshop on Human Motion, Austin, Texas, December 2000

Activity monitoring and summarization for an intelligent meeting room

Ivana Mikic, Kohsia Huang, Mohan Trivedi
Computer Vision and Robotics Research Laboratory
Department of Electrical and Computer Engineering
University of California, San Diego

Abstract

Intelligent meeting rooms should support efficient and effective interactions among its occupants. In this paper, we present our efforts toward building intelligent environments using multimodal sensor network of static cameras, active (pan/tilt/zoom) cameras and microphone arrays. Active cameras are used to capture details associated with interesting events. The goal is not only to make the system that supports multiperson interactions in the environment in real time, but also to have the system remember the past, enabling review of past events in an intuitive and efficient manner. In this paper, we present the system specifications and major components, integration framework, active network control procedures and experimental studies involving multiperson interactions in an intelligent meeting room environment.

1. Introduction

Intelligent environments are a very attractive domain of investigation due to both the exciting research challenges and the importance and breadth of possible applications. It is strongly influencing recent research in computer vision [1]. Realization of such spaces requires innovations not only in the computer vision [2, 3, 4, 5], but also in audio-speech processing and analysis [6, 7] and in the multimodal interactive systems area [8, 9, 10].

In this paper, we describe the system that handles multiperson interactions in an intelligent meeting room – Figure 1. It is being developed and evaluated in a multipurpose testbed called AVIARY (Audio-Video Interactive Appliances, Rooms and sYstems) that is equipped with four static and four active (pan/tilt/zoom) rectilinear cameras and two microphones.

2. Intelligent meeting room (IMR)

We consider IMRs to be spaces which support efficient and effective interactions among their human occupants. They can all be occupying the same physical space or they can be distributed at multiple/remote sites. The infrastructure which can be utilized for such intelligent rooms include a suite of multimodal sensory systems, displays, pointers, recording devices and appropriate computing and communications systems. The necessary "intelligence" of the system provides adaptability of the environment to the dynamic activities of the occupants in the most unobtrusive and natural manner.

Figure 1. An intelligent meeting room

The types of interactions in an intelligent environment impose requirements on the system that supports them. In an intelligent meeting room we identify three types of interactions:

- between active participants – people present in the room
- between the system and the remote participants
- between the system and the "future" participants

The first category of interactions defines the interesting events that the system should be able to recognize and capture. The active participants do not obtain any information from the system but cooperate with it, for example by speaking upon entering the room to facilitate accurate person identification.

Other two types of interactions are between the system and people that are not present in the room. Those people are the real users of the system. For the benefit of the remote participant, the video from active cameras that capture important details such as a face of the presenter or a view of the whiteboard should be captured and transmitted. Information on identities of active participants, snapshots of their faces and other information can be made available. The "future" participant, the person reviewing the meeting that happened in the past, requires a tool that graphically summarizes past events to easily grasp the spatiotemporal relationships between events and people that participated in them. Also an interface for interactive browsing and review of the meeting is desirable. It would provide easy access to stored information about the meeting such as identities and snapshots of participants and video from active cameras associated with specific events.

Interactions between active participants in a meeting room define interesting activities that the system should be able to recognize and capture. We identified three: a person located in front of the whiteboard, a lead presenter speaking and other participants speaking. A lead presenter is the person currently in front of the whiteboard. First activity should draw attention from one active camera that captures a view of the whiteboard. Other two activities draw attention from an active camera with the best view of the face for capturing the video of the face of the current speaker.

To recognize these activities, the system has to be aware of the identities of people, their locations, identity of the current speaker and the configuration of the room. Basic components of the system that enable described functionality are:

- 3D tracking of centroids using static cameras with highly overlapping fields of view
- Person identification (face recognition, voice recognition and integration of the two modalities)
- Event recognition for directing the attention of active cameras
- Best view camera selection for taking face snapshots and for focusing on the face of a current speaker
- Active camera control
- Graphical summarization/user interface component

Details associated with the overall architecture and specific components for IMR are in the next section.

3. The IMR components and system architecture

Integration of audio and video information is performed at two levels. First the results of face and voice recognition are integrated to achieve robust person identification. At a higher level, results of 3D tracking, voice recognition, person identification (which is itself achieved using multimodal information) and knowledge of the structure of the environment are used to recognize interesting events.

When a person enters the room, the system takes the snapshot of their face and sample of their speech to perform person identification using face and voice recognition [11, 12].

The system block diagram is shown in Figure 2. As mentioned before, it currently takes inputs from four static cameras with highly overlapping fields of view, four active cameras and two microphones. All of the eight cameras are calibrated with respect to the same world coordinate system using Tsai's algorithm [13].

Two PC computers are used. One performs 3D tracking of blob (people and objects) centroids based on input from four static cameras. Centroid, velocity and bounding cylinder information is sent to the other PC which handles all other system functions. For new people in the environment, the camera with the best view of the face is chosen and moved to take the snapshot of the face. The person is also required to speak at that time and the system combines face and voice recognition results for robust identification. Identity of the current speaker is constantly monitored and used to recognize interesting events together with 3D locations of people and objects and known structure of the environment. When such events are detected, the attention of active cameras is directed toward them.

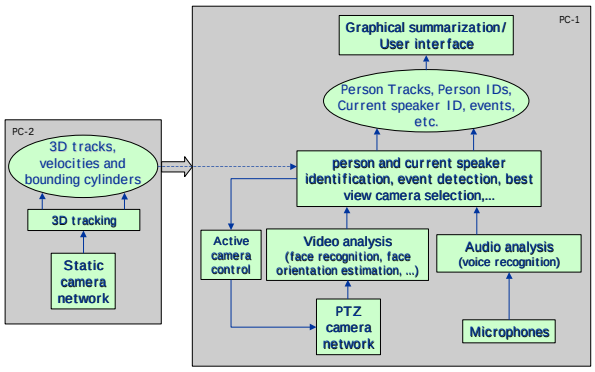

Figure 2. Block diagram of the system

3D centroid tracking. Segmentation results (object centroids and bounding boxes) from each of the four static cameras are used to track centroids of objects in the room and their bounding cylinders in 3D. Details of the tracking algorithm are given in [14]. The tracker is capable of tracking multiple objects simultaneously. It maintains a list of Kalman filters, one for each object in the scene. The tracker calculates updated and predicted

positions for each object in real time. Availability of up-to-date predictions allows feedback to the segmentation algorithm, which can increase its sensitivity in the areas where the objects are expected to appear. Figure 3 shows a typical input from four cameras with object centroids and bounding boxes calculated by the segmentation algorithm. Projections of the tracks back onto image planes and also projections onto the floor plane are shown.

Figure 3. 3D tracking – smaller crosshairs (barely visible) and green bounding boxes are segmentation results used by the tracker to compute 3D tracks and bounding cylinders. Larger crosshairs are projections of tracks back onto the image planes. Bottom: projections of the track onto the floor plane.

Person identification and current speaker recognition. Eigenface recognition algorithm [15] is currently utilized in the face recognition module. Human face is extracted from the snapshot image of camera network by skin color detection [16]. Face images of known people at certain facing angles are stored into the training face database. The face image is compared to the training faces in terms of distances in the eigenface space. The test face is then classified as a known person if the minimum distance to the corresponding training face is smaller then the recognition bound. For voice recognition, we use a text independent speaker identification module from the IBM ViaVoice SDK. When there is speech activity, clips of up to 5 seconds in length are recorded and sent to ViaVoice for recognition

The results of face and speaker recognition modules are fused together for robust person identification. Since ViaVoice does not provide access to confidence measures of recognition results, we are not able to make optimal decisions. Therefore, we perform the following fusing scheme. Each module gives output only if there is reasonable confidence associated with it. If only one module outputs a valid result, then it is taken as the final decision. If both modules output valid but different results, the output from face recognition is accepted if its confidence is above predetermined high value, otherwise the output from speaker recognition is accepted.

Event recognition for directing the attention of active cameras. This module constantly monitors for events described in the section 2. When a new track is detected in the room, it is classified as person or object depending on the dimensions of the bounding cylinder. This classification is used to permanently label each track. If classified as object, the camera closest to it takes the snapshot. If classified as person the camera with the best view of the face needs to be selected. The snapshot is then taken and person identification is performed. Each person track is labeled with person's name. Events are associated with tracks labeled as people (person located in front of a whiteboard, person in front of the whiteboard speaking and person located elsewhere speaking) and are easily detected using track locations and identity of the current speaker.

Best view camera selection. The best view camera for capturing the face is the one for which the angle between the direction the person is facing and the direction connecting the person and the camera is the smallest (Figure 4). Center of the face is taken to be 20cm from the top of the head (which is given by the height of the bounding cylinder).

There are three different situations where the best view camera selection is performed. First is taking snapshot of the face of the person that just entered the room. Second, if the person in front of the whiteboard is speaking a camera needs to focus on their face. The third situation is when the person not in front of the whiteboard speaks. In these three situations, we use different assumptions in estimating the direction the person is facing.

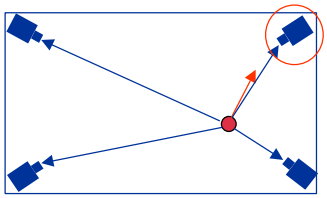

Figure 4. Best view camera is chosen to be the one the person is facing the most (maximum inner product between the direction the object is facing and direction toward a camera)

When a person walks into the room, we assume that they are facing the direction in which they are walking. If a

person is in front of a whiteboard (location of which is known), one camera focuses on the whiteboard (Figure 5). If the person starts speaking, a best view camera needs to be chosen from the remaining cameras to focus on that person's face. Since the zoomed-in whiteboard image contains person's head, we use that image to estimate the direction the person is facing. Due to the hairline, the ellipse fitted to the skin pixels changes orientation as person turns from far left to far right (Figure 6). We use skin detection algorithm described in [16]. If skin pixels are regarded as samples from a 2D Gaussian distribution, the eigenvector corresponding to the larger eigenvalue of the 2×2 covariance matrix describes the orientation of the ellipse. A lookup table based on a set of training examples (Figure 7) is used to determine the approximate angle between the direction the person is facing and the direction connecting the person and the camera that the whiteboard image was taken with. These angles are not very accurate, but we have found that this algorithm works quite reliably for purposes of best view camera selection.

Figure 5. Person close to the whiteboard draws attention from one active camera

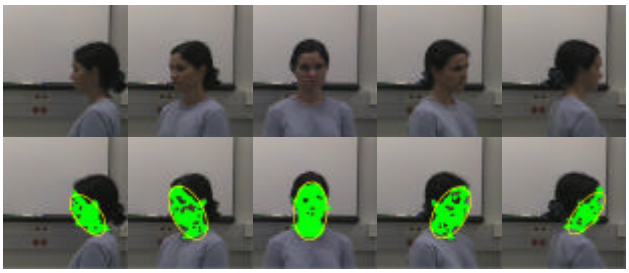

Figure 6. Face orientation estimation for best view camera selection

In the third case, where person elsewhere in the room is speaking, we assume they are facing the person in front of the whiteboard if one is present there. Otherwise, we assume they are facing the opposite side of the room. The first image obtained with the chosen camera is processed using the algorithm described in the previous paragraph and the camera selection is modified if necessary.

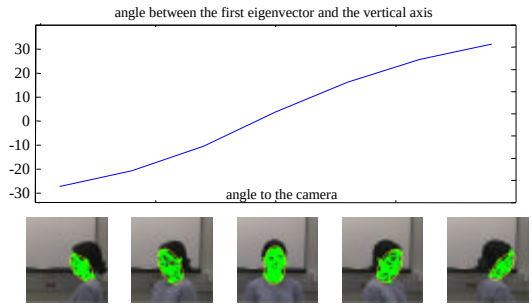

Figure 7. Lookup table for face orientation estimation computed by averaging across training examples

Active camera control. Pan and tilt angles needed to bring the point at a known location to the center of the image can be easily computed using the calibrated camera parameters. However, the zoom center usually does not coincide with the image center. Therefore, the pan and tilt angles needed to direct the camera toward the desired location have to be corrected by the pan and tilt angles between the center of the image and the zoom center. Otherwise, for large magnifications, the object of interest may completely disappear from view.

A lookup table (Figure 8) is used to select a zoom needed to properly magnify the object of interest (person's face or a whiteboard). Zoom magnification is calibrated using the model and the algorithm described in [17]:

$$x' = M(n)[x - C_x] + C_x$$
$$y' = M(n)[y - C_y] + C_y \qquad (1)$$

where n is the current zoom value, $M(n)$ is the magnification, C_x and C_y are the coordinates of the center of expansion or the zoom center, x and y are the coordinates of a point at zero zoom and x' and y' are coordinates of the same point at zoom n.

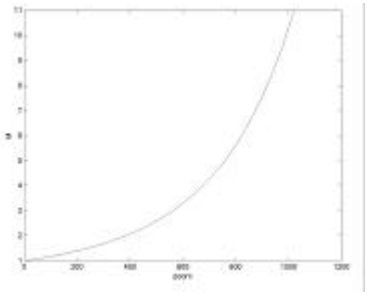

Figure 8. Magnification for different zoom values for on of the active cameras

Magnifications are computed for a subset of possible zoom values defined by a chosen zoom step. Magnifications for other zoom values are interpolated from the computed ones. The magnifications are obtained using a slightly modified version of [17]. Two images taken with two different zoom values are compared by

shrinking the one taken with the larger zoom using the Equation 1. The value of magnification (will be smaller than 1) that achieves best match between the two images is taken to be the inverse of the magnification between the two images. The algorithm described in [17] was written for outdoor cameras where objects present in the scene are more distant from the camera than in the indoor environments. Therefore, instead of comparing images at different zooms to the one taken at zero zoom as done in [17], we always compare two images that are one zoom step apart. The absolute magnification for a certain zoom value with respect to zero zoom is computed by multiplying magnifications for smaller zoom steps (Figure 8). However, we could not reliably determine the location of the zoom center using this algorithm. Instead, we determine its coordinates manually by overlaying a crosshair over the view from the camera and zooming in and out until we find a point that does not move under the crosshair during zooming.

Graphical summarization/user interface. The history is summarized graphically for easy review and browsing of information the system collected about the environment. The 3D graphic representation shows the room floor plan and the third axis represents time (**Figure 9**). Tracks are color-coded and represented by one shape (i.e. sphere) when the person is not speaking and by a different one (i.e. cube) when the person is speaking. The floorplan shows important regions like whiteboard and doors. This graphical representation effectively summarizes events that system can detect and trajectories and identities of people involved. It also serves as a user interface. By clicking on a colored shape, the user is shown the face snapshot and the name of the person associated with the track and the video associated with the event the shape corresponds to can be replayed.

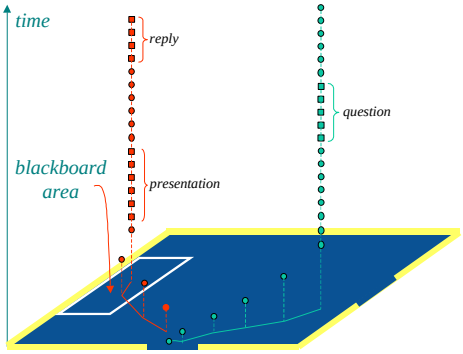

Figure 9. Graphical summarization of the events in the environment. A presenter (red) and another participant (green) were present.

4. System performance

The described system is operating quite reliably. In [14], we have described the experiments on the accuracy of centroid tracking and have reported good results with maximum errors around 200mm. We currently have only five people in the face and speaker databases, so the person identification accuracy based on both modalities is practically 100%. Also, recognition of the current speaker performs with nearly perfect accuracy if silence in a speech clip is less then 20% and clip is longer than 3 seconds. The results are very good for clips with low silence percentage even for shorter clips, but gets erroneous when silence is more than 50% of the clip. However, there is a delay of 1-5 seconds between the beginning of speech and the recognition of the speaker, which causes a delay in recognizing activities that are concerned with the identity of the current speaker.

If the person faces the direction they are walking, the camera selection for acquisition of face snapshots also works with perfect accuracy. It would, of course, fail if person turned their head while walking. The camera selection for focusing on the face of the person that is talking in front of the whiteboard succeeds around 85% of the time. In the case of the person talking elsewhere in the room, our assumption that they are facing the person in front of the whiteboard or the opposite side of the room is almost always true. This is due to the room setup – there is one large desk in the middle of the room and people sit around it – therefore almost always facing the opposite side of the room, unless they are talking to the presenter

We can store all information needed to access appropriate parts of the video that correspond to the events the user selects from the interface. From the interface, the user can view identities and face snapshots of people associated with different tracks by clicking on the corresponding colored shape. For remote viewing, the videos from active cameras that capture interesting events can be transmitted together with the other information needed to constantly update the summarization graph. See Figure 10 for the illustration of the system operation.

5. Concluding remarks

We have presented our investigations toward building the multimodal intelligent environments that provide awareness of people and events at several resolution levels: from a graphical summarization of the past and ongoing events to the active camera focus on interesting events and people involved in them.

Next step in this investigation would be more detailed and sophisticated audio and video analysis that would use high-resolution information collected by the active camera and microphone network. This would include posture estimation, gesture recognition, speech recognition, lip-reading, etc.

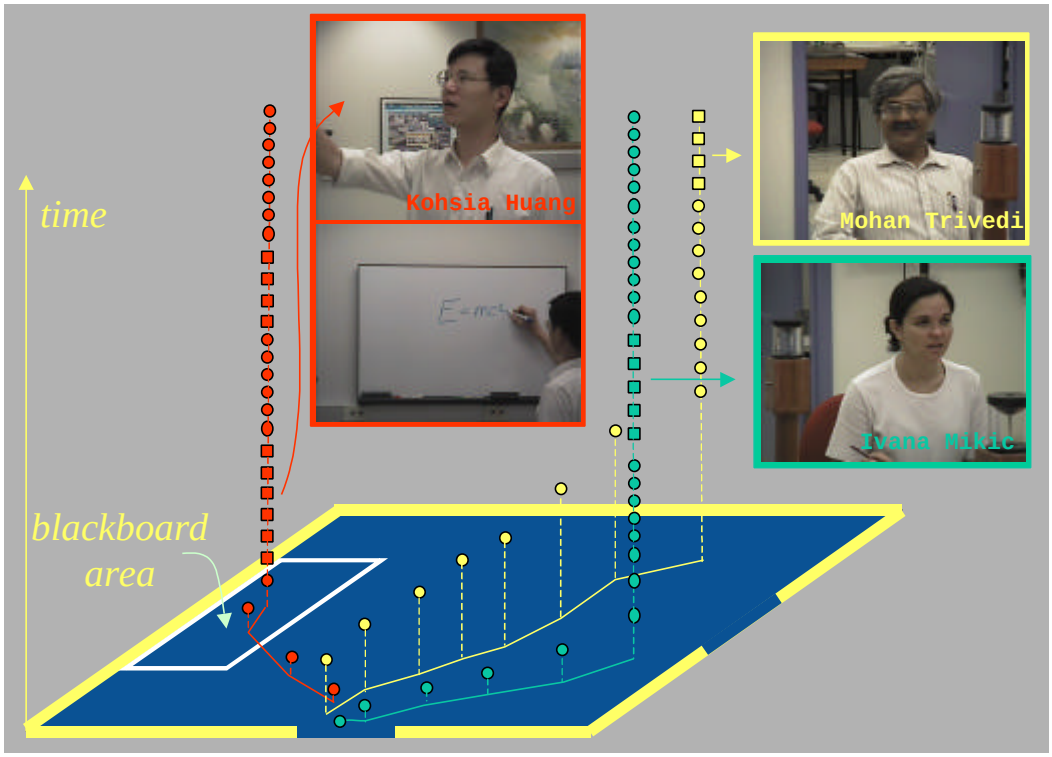

Figure 10. Illustration of the system operation. Interesting activities attract attention from active cameras. That video can be transmitted to remote viewers or stored for later reviewing. Every "object" in this graphical summarization is associated with information needed to access the appropriate portion of video, face snapshots and identity information

References

[1] A. Pentland, "Looking at People: Sensing for Ubiquitous and Wearable Computing", *IEEE. Trans. PAMI*, 22(1), Jan 2000, pp. 107-119

[2] D. Gavrila, "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding*, 73(1), Jan 1999, pp. 82-98

[3] V. Pavlovic, R. Sharma, T. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review", *IEEE Trans. PAMI*, 19(7), July 1997, pp. 677-695

[4] R. Cipolla, A. Pentland (editors*), Computer Vision for Human-Machine Interaction*, Cambridge University Press, Cambridge, UK, 1998

[5] R. Chellappa, C. Wilson, S. Sirohev, "Human and Machine Recognition of Faces: A Survey", *Proc. IEEE*, 83(5), pp. 705-740, 1995.

[6] L. Rabiner, B. Juang, "Fundamentals of Speech Recognition", Englewood Cliffs, NJ: Prentice-Hall 1993.

[7] M. Brandstein, J. Adcock, H. Silverman, "A closed-form location estimation for use with room environment microphone arrays", *IEEE Trans. Speech and Audio Processing*, 5(1), Jan. 1997, pp. 45-50

[8] R. Sharma, V. Pavlovic, T. Huang, "Toward Multimodal Human-Computer Interface", *Proc. IEEE*, 86(1), May 1998, pp. 853-869

[9] M. Trivedi, B. Rao, K. Ng, "Camera Networks and Microphone Arrays for Video Conferencing Applications", *Proc. Multimedia Systems Conf.*, Sep 1999

[10] C. Wang, M. Brandstein, "A hybrid real-time face tracking system", Proc. IEEE ICASSP '98, p.3737-40

[11] M. Trivedi, I. Mikic, S. Bhonsle, "Active Camera Networks and Semantic Event Databases for Intelligent Environments", *IEEE Workshop on Human Modeling, Analysis and Synthesis*, June 2000

[12] M. Trivedi, K. Huang, I. Mikic, "Intelligent Environments and Active Camera Networks", *IEEE Conf. SMC 2000*

[13] R. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses", *IEEE J. Robotics and Automation*, RA-3(4), 1987

[14] I. Mikic, S. Santini, R. Jain, "Tracking Objects in 3D using Multiple Camera Views", *Proc. ACCV2000*, Jan 2000, pp. 234-239

[15] M. Turk, A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Conf. Comput. Vis and Patt. Recog., Maui, HI, USA*, pp. 586-591, Jan. 1991.

[16] J. Yang, A. Waibel, "A Real-Time Face Tracker," *Proc. WACV'96, Sarasota, FL, USA*, 1996.

[17] Collins, Tsin, "Calibration of an Outdoor Active Camera System", CVPR '99, Fort Collins, CO, June 1999, pp. 528-534