Tracking Multiple People with Multiple Cameras

Scott Stillman, Rawesak Tanawongsuwan, and Irfan Essa

College of Computing, GVU Center,
Georgia Institute of Technology
Atlanta, GA 30332-0280 USA
{*sstil,tee,irfan*}*@cc.gatech.edu*
www.gvu.gatech.edu/perception/projects/smartspaces/

Abstract

In this paper we present a robust real-time method for tracking multiple people from multiple cameras. Our method uses both static and Pan-Tilt-Zoom (PTZ) cameras. The static cameras are used to locate people in the scene, while the PTZ cameras "lock-on" to the individuals and provide visual attention. The system provides consistency in tracking between PTZ cameras and works reliably well when people occlude each other. The underlying visual processes rely on color segmentation, movement tracking and shape information to locate target candidates and color indexing methods to register these candidates with the PTZ cameras.

1. Introduction

One of the goals of building an intelligent environment is to make it more aware of the user's presence so that the interface can *seek out* and serve the user [1]. This work addresses the ability of a system to determine the presence of humans and track their locations in indoor environments (*e.g.,* business offices, class rooms, smart rooms, etc.).

A problem that persists in static camera systems that track people is that these cameras cannot provide a useful image resolution for further perceptual analysis (*e.g.,* face recognition and expression recognition). Another problem is the inability of the vision system to handle occlusions. One method for dealing with these problems is to have a number of cameras available to foveate on the targets of interest. Our system addresses both of these issues.

The two main avenues we are pursuing to achieve robust tracking of multiple people from multiple cameras are:

Locating people: Determine the location of people in a calibrated scene by color tracking. Using this method the static cameras locate people in the scene and their positions are sent to the PTZ cameras.

Visual attention: Develop a mechanism for visual attention such that the PTZ cameras "lock" on to each of the users and zoom in to acquire high resolution imagery of people's faces and expressions. The PTZ and static cameras track people and handle occlusions between people as they move around in real-time.

In the rest of the paper, we briefly discuss the related work in the area of tracking people followed by a technical discussion of our approach. We conclude by presenting experimental results of our work and possible future extensions.

2. Related Work

Many real-time systems have been developed for face and gesture tracking each varying in function and detail [3, 4, 6, 10, 11, 12, 14, 16, 17]. Most of these methods use a combination of skin-color segmentation, motion analysis, and feature detection to locate faces. Here we discuss only a few of these that are closely related to our work.

Crowley and Berard [6] present an architecture for real-time systems where a supervisory controller selects and activates various visual processes in order to ascertain a degree of confidence concerning observed data. These visual processes are blink detection, color histogram matching, and normalized cross correlation. Blink detection in particular requires a high degree of resolution from the input image and appears to be a limiting factor in their system.

Goodridge and Kay [10] present a system that uses a camera with a wide angle lens to gather a coarse representation of the environment and uses two microphones to perform auditory localization for pin-pointing the user location. This data is combined to control a single pan-tilt-zoom camera. Other significant work on foveating systems has been undertaken by [2, 5, 8].

Yang and Waibel [14] and Wang and Chang [16] focus on tracking with a single camera and both use skin-color segmentation as the first level processing step.

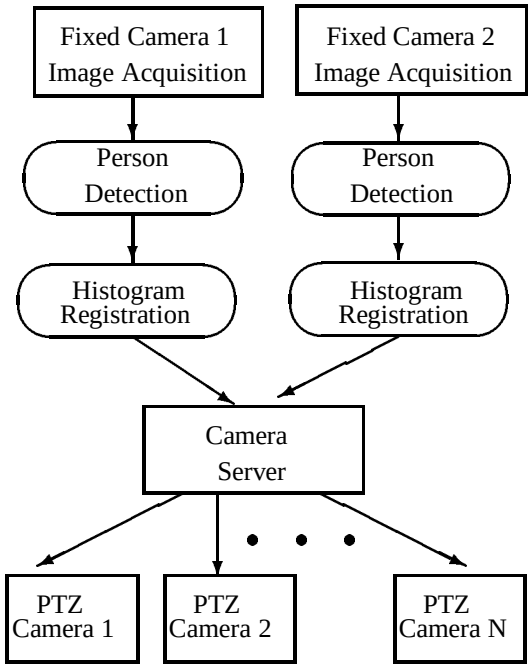

Figure 1. *The general system architecture could contain a number of PTZ cameras.*

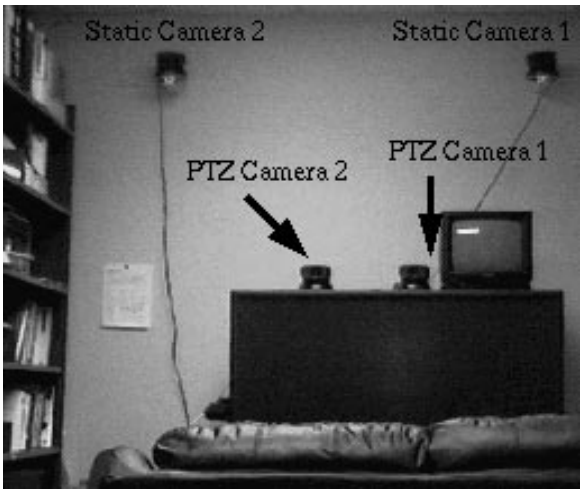

Figure 2. *The camera configuration consists of two static cameras that are mounted at the top of the picture and two PTZ cameras on top of the big screen television.*

Our system draws from parts of all the above mentioned systems to provide tracking for each static camera and uses a triangulation method to locate users once they are registered through a color-indexing method [13]. It also extends the above methods to deal with multiple people.

3. Multiple Cameras, Multiple People

We are interested in developing an architecture to track multiple people in complex scenes using multiple static and PTZ cameras. Figure 1 shows a general system configuration of two visual streams from static cameras feeding into the camera server, resulting in active tracking by PTZ cameras. At present we are working on a system for tracking that uses two static and two PTZ cameras. Figure 2 shows a typical layout of the two static cameras and two PTZ cameras in the lab. Such a set-up is possible in any scenario, provided we can calibrate the space.

The static cameras send their data to different processes where face detection and local registration between frames occur. The target locations of faces in each image plane are sent to the camera server for registration between the static cameras. The camera server uses triangulation of the target locations acquired using skin-color segmentation to obtain world coordinates of the people being tracked (Figure 3). The world coordinates of each target are used to update the pan angle and zoom factor for each PTZ camera. If a person becomes occluded or leaves the scene, the PTZ camera remains at the last updated location until the person re-appears.

Currently, our system works for only two people in the viewing area. We also require that these two people are not wearing the same color of shirt or blouse. Swain and Ballard's algorithm [13] is used in the histogram registration process which forces us to restrict changes in the illumination intensity.

In the following sections we describe the methodology and implementation issues in greater detail. The person detection and histogram registration sections describe the processing performed on a single data stream from each static camera, and the camera server section describes the integration of the data from these cameras.

3.1 Person Detection

Skin-color segmentation is a key step in many existing real-time people tracking systems. Our system uses skin-color segmentation based on the work of Yang and Waibel [14] to get an initial set of candidate faces. Yang, Lu, and Waibel [15] have shown that human skin colors cluster in a small region in a normalized color space and that under certain lighting conditions, the skin-color distribution can be characterized by a multivariate normal Gaussian distribution.

The image pixels from the static cameras are projected into a normalized color space defined by

$$g_r = \frac{g_R}{g_R + g_G + g_B}, \quad (1)$$

$$g_g = \frac{g_G}{g_R + g_G + g_B}. \quad (2)$$

A pixel g is represented by its red (g_R), green (g_G), and blue (g_B) components in RGB space. g_r and g_g represents the projected components of the pixel into this normalized space.

The Gaussian model parameters (means and variances) are obtained by processing skin-tone samples captured off-line. Next, a Gaussian distribution is created from these parameters and is used to generate a Look-Up Table (LUT) of values to be stored during system initialization. The LUT is used to accept or reject pixels as skin pixels based on the pixel's proximity to the Gaussian peak.

Next, salt-and-pepper noise is removed from the image frame by using a median-filtering technique. The resulting image $I(x, y)$ from this process is smoothed and a horizontal projection of the skin pixels in the image is calculated. The coordinate system of the image frame is expressed with x in the horizontal direction and y in the vertical direction.

$$Proj(x) = \sum_{y=0}^{height-1} I(x, y) \quad x = 0, 1, \ldots width-1 \quad (3)$$

$Proj(x)$ looks like a range of mountains with peaks and valleys. The region surrounding each mountain peak is likely to contain a face. The peaks and their surrounding areas of $Proj(x)$ above some threshold are extracted to produce a set of vertical slices Θ of the image. Next, a vertical projection is created from these slices to determine the face candidate regions.

$$Proj(y) = \sum_{x \in \Theta} I(x, y) \quad y = 0, 1, \ldots height - 1 \quad (4)$$

Extraction of the peaks and surrounding areas of $Proj(y)$ results in an initial set of candidate faces. Using heuristics for the size and the aspect-ratio of human faces[1], many of the non-face candidates can be eliminated.

The two static cameras perform inter-frame registration of the people based on color, size, velocity, and aspect ratio attributes (with the assumption that they do not change faster than the video capture frame-rate). A scoring method is used to combine this data to produce a single numerical score of the likelihood that a candidate region is indeed a face. If this number is greater than a preset threshold, it is declared a face object and is added to a list of face objects.

After the system is initialized and a face has been detected, a rectangular region or swatch is automatically cropped below the face object and is used to generate a color histogram which is stored in the face object's data structure. This is based on the assumption that a person is standing upright, which results in this swatch extracting the information about the person's clothes. The object's histograms is used to register and provide tracking consistency. The registration process is described in detail below.

[1]The golden ratio (artistic-ese) is used as the comparison aspect ratio and is defined as: $\frac{height}{width} \equiv \frac{1+\sqrt{5}}{2}$.

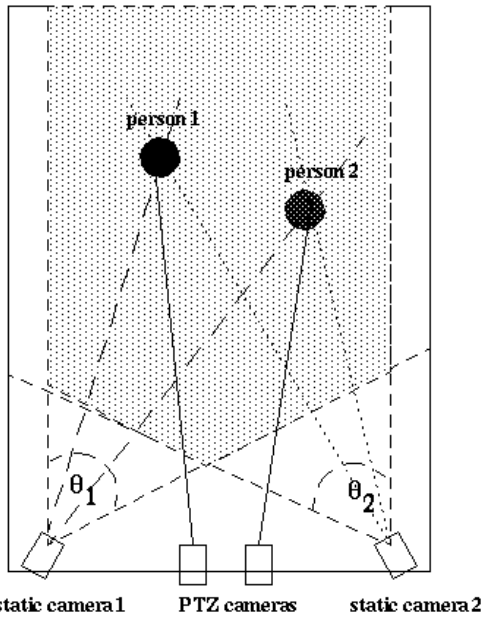

Figure 3. *A bird's eye view of the Field of View (FOV) angles used for triangulation.*

3.2 Histogram Registration

A registration process of people between the static cameras is needed to achieve consistent tracking with the PTZ cameras and for calculating the locations of people. Registration is performed in the Histogram Registration module for each stream and an identification number is assigned to each face object before the face object list is sent to the camera server shown in Figure 1.

The registration process is first initialized by getting swatch samples from each person. The system then creates color histograms from those swatches that are later used to register people as they move about the scene. The success of this registration method relies on the differences of cloth colors.

Color histograms are constructed by counting the pixels of a given value in an image. The color space used consist of three axes B-G, G-R, and R+G+B. Each axis is coarsely partitioned into 16 bins. The time to construct a histogram is proportional to the number of image pixels in the swatch.

During initialization, model histograms M_1 and M_2 are created from the area below face regions of person 1 and

Figure 4. *An earlier implementation of our system tracking and identifying 3 people in a scene using only two static cameras. People are recognized by the color of their clothes. Note that in such images face recognition is very hard as each face image has a resolution of about 25x40 on average (the image size is 320x240).*

person 2. After that, in each frame, the person detection module outputs some number of candidate regions K. Candidate histograms C_1, C_2, ..., C_K are constructed in the same way as model histograms. The histogram intersection then is computed between the model histogram M_1 and a list of candidate histograms.

$$score_m = \frac{\sum_{i=1}^{N} min(M_1(i), C_m(i))}{\sum_{i=1}^{N} M_1(i)} \quad m = 0, 1, \ldots K \tag{5}$$

The best match for person 1 is the maximum score that is above some threshold T. We do the same for the model histogram of the second person M_2 to get the best match. Once the histogram registration step is complete, the two most likely regions are then identified as person 1 and person 2. The histogram module then sends the (x, y) centroid coordinates in the image space of these two regions and their respective person identifiers to the server.

3.3 The Camera Server

The camera server accepts processed data in the form of a face object list from the two static cameras. The server process waits until it receives a list from both cameras before it begins the location procedure. Next the server uses the person identification numbers of each face object from each static camera to register the faces between cameras. Simple geometry is used to triangulate the face objects to obtain world coordinates of the face object in two dimensions (x, y). Figure 3 shows the Field of View (FOV) angles used for triangulation. The calibration of the space requires measuring the base distance between the static cameras and calculating the angle of the frustum. Camera 1 is labeled as the origin in world coordinates and the other cameras position are measured relative to the origin. The height dimension can be obtained from the relative positions of faces in the static camera views.

The coordinates determined from the above process are sent to another process that controls the pan-tilt-zoom of the PTZ cameras. This process accepts the coordinates and sends motor commands to its corresponding PTZ camera.

4. Experimental Results

Figure 4 shows the output of an earlier implementation of our system before we added the PTZ cameras. The resolution of the cropped face alone in most cases was not adequate for further analysis (*e.g.,* face recognition, gaze tracking, or expression recognition). In this case, we used the color information of the swatch to identify the people.

In our current system, we incorporate PTZ cameras allowing the system to foveate to objects of interest in the scene. Such visual attention results in an increase in resolution of the object being tracked (in our case the faces of the users). Additionally we use a histogram matching technique to achieve tracking consistency. This method makes it possible to reliably track multiple people even when they occlude one another. Figure 5 shows the initialization step of our system. The two people must first face the camera for a good swatch to be extracted. This is done automatically when the system is started. After that point, each PTZ camera is assigned to a person based on which part of the room they are in. Figure 6 (A) and (B) shows the view of each of the static cameras with the faces and swatches marked. The swatch area is represented by a white bounding box under the face.

Once the initialization step is complete, the PTZ cameras follow their registered guest as they move about the room. Figure 7 shows the view of one of the static cameras when the two people cross paths. And finally, Figure 6 (C)

(A) *(B)* *(C)* *(D)*

Figure 6. *(A) & (B) Views from two static cameras showing the result from the person detection module. (C) & (D) Views from two PTZ cameras. We use the results of the triangulation process to decide the zoom scale factor (*i.e.*, the distance from a person to the PTZ camera).*

Figure 5. *The system is initialized by creating swatch histograms of two people once they are present in the scene.*

Figure 7. *This is a view from one of the static cameras when two people cross paths. Tracking continues despite occlusions.*

and (D) shows the views from the two PTZ cameras.

Our system at present runs in real-time using two Silicon Graphics (SGI Indy R5000) workstations. We use SONY EVI-D30 PTZ cameras. The PTZ cameras are controlled through their serial ports.

5. Conclusions & Future Work

In this paper we present an approach that uses skin-color segmentation and color indexing methods to locate multiple people in a calibrated space. By combining multiple static cameras and PTZ cameras we have developed an architecture that provides higher resolution imagery of faces for face processing applications. The tracking methods we have proposed work well under simple occlusions that occur when people cross one another in the scene. We demonstrate a real-time system that tracks two people reliably using two static and two PTZ cameras.

We are working on several extensions to improve the the system's ability to detect faces and to increase the accuracy of the tracking and registration processes. A few of these extensions are to:

- employ face specific visual processes to verify the existence of faces in the image,
- use real-time stereo vision to acquire 3D position of the users,
- apply Kalman filtering techniques to improve the tracking by the PTZ cameras,
- incorporate audio information to develop phased-array microphones for auditory localization, AND
- experiment with this system in more complex scenarios like classrooms and meeting rooms.

Acknowledgment

The authors would like to thank Q. Alex Zhao and Xin Zhong for helping with the initial implementation of this system and for their programming expertise.

References

[1] G. Abowd, C. Atkeson and I. Essa, "Potential Applications of Perception in Future Computing Environments," *Workshop on Perceputual User Interfaces*, Alberta, Canada, October 1997.

[2] S. Basu, *et. al.*, "Vision-Steered Audio for Interactive Environments," *Proceedings of IMAGE'COM 1996*, May 1996.

[3] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

[4] A. Bobick, J. Davis and S. Intille, "The KidsRoom: An example application using a deep perceptual interface," *Workshop on Perceputual User Interfaces*, Alberta, Canada, October 1997.

[5] M. Colloberti, *et. al.*, "LISTEN: A System for Locating and Tracking Individual Speakers," ICAFGR96, 1996.

[6] J. Crowley and F. Berard, "Multi-Modal Tracking of Faces for Video Communications," *IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997.

[7] J. Crowley and Y. Demazeau, "Principles and techniques for sensor data fusion," *Multisensor Fusion for Computer Vision*. NATO ASI Series, Springer-Verlag, pages 5–27, 1993.

[8] T. Darrell, I. Essa and A. Pentland, "Attention-driven Expression and Gesture Analysis in an Interactive Environment," IWAFGR, pages 135–140, 1995.

[9] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. Wiley, N.Y., 1973.

[10] S. Goodridge and M. Kay, "Multimedia Sensor Fusion for Intelligent Camera Control," *Proc. of 1996 IEEE/SICE/RSJ Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, pages 655–662, NJ, December 1996.

[11] H. Graf, *et. al.*, "Multi-Modal System for Locating Heads and Faces," *Int. Conf. on Automatic Face and Gesture Recognition*, pages 88–93, Vermont, October 1996.

[12] T. Darrell, *et. al.*, "Integrated person tracking using stereo, color, and pattern detection." *Computer Vision and Pattern Recognition '98*, pages 601–608, Santa Barbara, CA, June 1998.

[13] M. Swain and D. Ballard, "Color Indexing," *Int. J. of Computer Vision*, 7(1):11–32, 1991.

[14] J. Yang and A. Waibel, "A Real Time Face Tracker," *IEEE Workshop Appl. Comput. Vision*, pages 142–147, Los Alamitos, CA USA 1996

[15] J. Yang, W. Lu and A. Waibel, "Skin-Color Modeling and Adaptation," *Proc. of ACCV'98*, 2:687–694, Hong Kong, 1998.

[16] H. Wang and S. Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video," *IEEE Transactions on Circuits and Systems for Video Technology*, 7(4):615–628, August 1997.

[17] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.