

5-2004

CAMEO: Camera Assisted Meeting Event Observer

Paul E. Rybski
Carnegie Mellon University

Fernando de la Torre
Carnegie Mellon University

Raju Patil
Carnegie Mellon University

Carlos Vallespi
Carnegie Mellon University

Manuela M. Veloso
Carnegie Mellon University

See next page for additional authors

Follow this and additional works at: <http://repository.cmu.edu/robotics>

 Part of the [Robotics Commons](#)

Published In

IEEE International Conference on Robotics and Automation, 2004. Proceedings, 2, 1634-1639.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Robotics Institute by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Authors

Paul E. Rybski, Fernando de la Torre, Raju Patil, Carlos Vallespi, Manuela M. Veloso, and Brett Browning

CAMEO: Camera Assisted Meeting Event Observer

Paul E. Rybski, Fernando de la Torre, Raju Patil,
Carlos Vallespi, Manuela M. Veloso, Brett Browning
School of Computer Science
Carnegie Mellon University

5000 Forbes Ave., Pittsburgh, PA, 15213

Email: {prybski,ftorre,raju,cvalles,veloso,brettb}@cs.cmu.edu

Abstract—Static cameras are pervasive in a variety of environments. However it remains a challenging problem to extract and reason about high-level features from real-time and continuous observation of an environment. In this paper, we present CAMEO, the Camera Assisted Meeting Event Observer, which is a physical awareness system designed for use by an agent-based electronic assistant. CAMEO is an inexpensive high-resolution omnidirectional vision system designed to be used in meeting environments. The multiple camera design achieves the desired high image resolution and lower cost that can be achieved when compared to traditional omnicones that make use of a single camera and mirror solution.

I. INTRODUCTION

Transcribing events from a formal meeting setting (such as who spoke and what was discussed) into a digital form that can be searched and analyzed is a fairly tedious task for a human to do. Systems such as optical character recognition (OCR) and voice-to-text converters speed up the transfer of analog (written and spoken) data into a digital form that can be more easily indexed and manipulated. However, most of these tools are relegated to passive data entry roles where relevant information has already been identified so that it can be scanned or read aloud into a machine. Automatic transcription agents in the near future will be far more interactive and will be able to listen to groups of people talking at a meeting and be able to record (with the permission of the attendees) the spoken, written, and gestured communication. Afterwards, this information can be automatically collated and sorted such that it can easily answer questions such as "What was the third bullet on slide 15?", or "What was the action item decided on while I was out of the room?" Such a personal digital assistant promises to be able to give the user the ability to recall events throughout the working day whose importance might not have been realized at the time and as such would never have been manually recorded.

In order to effectively interact with humans in a natural fashion, such an intelligent agent must have a robust physical awareness system with which it can sense humans as they perform tasks in the real world. The Camera Assisted Meeting Event Observer (CAMEO), shown in Figure 1 is a sensory system designed to provide an electronic agent with physical awareness of the real world. CAMEO consists of a set of five cameras oriented in such a way as to capture a panoramic video stream of the world. This stream is scanned for human activity by identifying the positions of human faces found

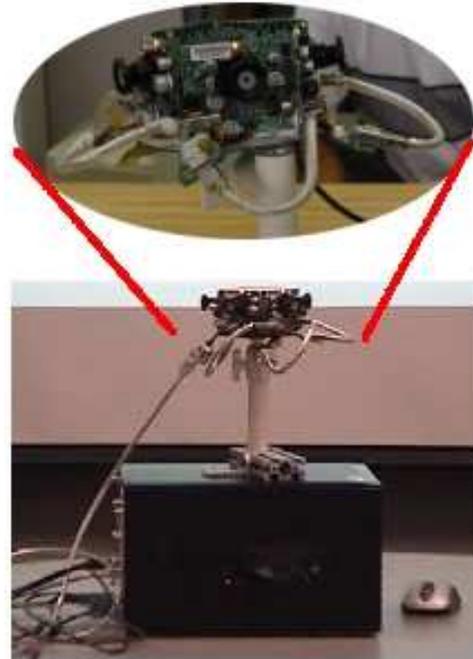


Fig. 1. The complete CAMEO system consists of five calibrated firewire cameras and a portable image-processing workstation that can be connected to other systems for doing logging and behavior recognition (not shown).

in the image. All of the extracted face information, and the context in which it is found, is logged for future viewing, post-processing, and modelling. Instead of instrumenting meeting rooms with large numbers of calibrated cameras, CAMEO is intended to be used more like a speaker phone for a conference call. That is, a CAMEO device will be brought into a meeting and simply placed in the center of the room without requiring special calibration. As such, CAMEO is designed to be used in environments where those who are participating in the meetings agree to and welcome the use of such an electronic assistant.

CAMEO is part of a larger effort to develop an enduring personalized cognitive assistant that is capable of helping humans handle the many daily business/personal activities that they engage in. This larger project¹, called CALO (Cognitive Agent that Learns and Organizes), is developing a personalized

¹Funded by DARPA and managed by SRI International

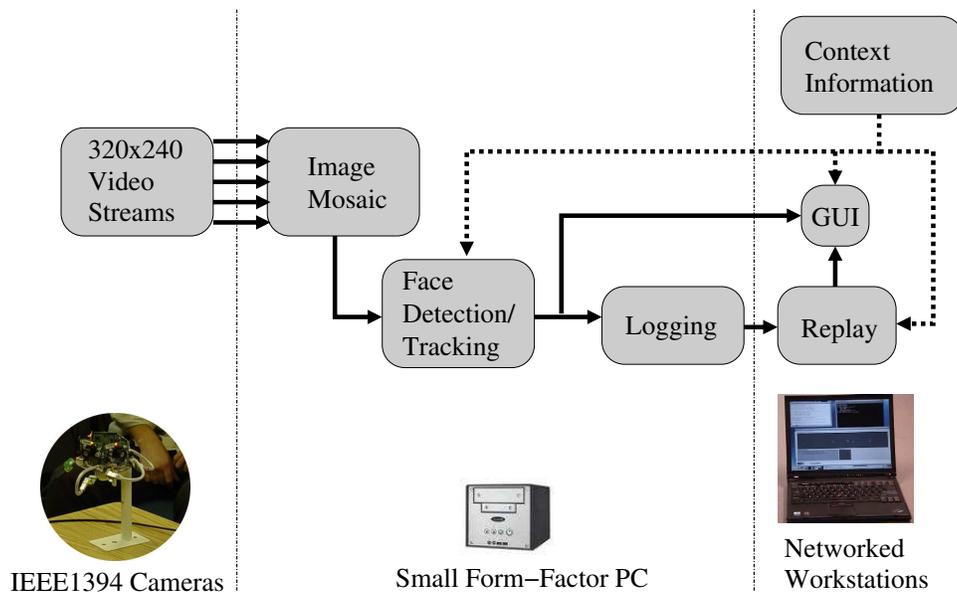


Fig. 2. Schematic diagram of the CAMEO system. The five firewire (IEEE1394) cameras capture individual data streams that are merged into a panoramic mosaic by the PC. This image mosaic is passed into the face detector/tracking module for processing. All extracted faces are logged (along with the video stream) and this information is passed onto a networked workstation for replay purposes as well as for live display. The dashed lines show conceptually how high-level person/meeting information would be input into the system by other knowledge bases (which are beyond the scope of CAMEO).

omnipresent computational resource that will be able to handle routine tasks/events, anticipate predictable user needs and prepare for them appropriately, and assist the user in handling unexpected events. The learning component of CALO will observe the users' physical and electronic activities and will recognize and classify patterns of activities. Once a set of activities is classified, CALO will adapt to changing user preferences/requirements as well as anticipate its user's future needs. A detailed discussion of the full extent of the CALO project is beyond the scope of this paper. Thus, only CALO's physical awareness effort, as instantiated in the CAMEO system, will be described.

II. THE CAMEO SYSTEM

The CAMEO system consists of five firewire cameras mounted in a circle so as to provide an omnidirectional view of the world. The five cameras are daisy-chained on their bus which allows them to all be plugged into a single interface card on the controlling PC. A Small Form-Factor (SFF) PC equipped with a 3.0GHz Pentium 4 processor captures the images and handles all of the image processing algorithms.

A conceptual diagram of CAMEO is shown in Figure 2. Once captured, the five raw video data streams are merged into a consistent panoramic image mosaic. This mosaiced image is passed into the face detection module that returns the (x, y) (image coordinates) positions of all of the faces in the image. This data is broadcast to other computers over TCP sockets and is also saved to a log file. Other machines running a GUI can observe the data returned from CAMEO in real time as well as replay the log files offline. Since CAMEO is designed to link into the larger CALO system, it would have access to a knowledge base that would provide information regarding the

meeting context. In the figure, these connections are denoted by dashed lines.

Traditionally, omnidirectional images have been obtained by omnicaamera systems consisting of a single camera aimed at a curved mirror [1][2]. The primary disadvantage of using an omnicaamera is that the entire panoramic image is projected onto a single CCD/CMOS imaging surface. In addition, data at the center of the imaging surface is typically unusable because the mirror reflects the lens of the camera itself. Compounding this resolution loss is the fact that some of the areas of interest are compressed and need to be dewarped to a Cartesian projection before they can be used for image analysis. By using multiple cameras, all of the data returned from each camera can be used. The downside of using five cameras is the fairly large amount of data that must be transferred over the firewire bus. However, by capturing images at 320x240 pixels at 30 fps, only half of the total bandwidth of the firewire interface is used. Additionally, each of the firewire cameras is roughly under \$100 each and so CAMEO is much less expensive than an omnicaamera which typically requires a precision ground mirror that can be very expensive to produce and time consuming to calibrate/align properly.

III. MOSAIC GENERATION

To get a global description of the scene, CAMEO integrates the images coming from all the cameras into a single mosaic. Because the cameras do not share a common center of projection, parallax effects occur due to the translational component between the cameras. With translational displacement between cameras, the geometric transformation that relates two images becomes depth dependent (the parallax effect becomes more evident at shorter distances). One possible solution involves



Fig. 4. a) Original images. b) Mosaic image.

computing depth for each point [3], however, this approach will be very expensive for real time applications. In our case, we assume that the people are one order of magnitude ($\sim 2m$) further than the biggest translation between the optical camera's center.

We use lenses with small focal length (2.5 mm) to obtain wide field of view (110°), however, this fact introduces a huge radial distortion. We calibrate the camera to get its intrinsic parameters (focal length f_x, f_y , and principal point x_o, y_o) and the parameters to correct the radial distortion (k_1, k_2, k_3, k_4). We use the following radial distortion model:

$$\begin{aligned} x_n &= \frac{X}{Z} \\ y_n &= \frac{Y}{Z} \\ r^2 &= x_n^2 + y_n^2 \\ u_p &= (1 + k_1 r^2 + k_2 r^4)x_n + 2k_3 x_n y_n + k_4(r^2 + 2x_n^2) \\ v_p &= (1 + k_1 r^2 + k_2 r^4)y_n + 2k_3 x_n y_n + k_4(r^2 + 2y_n^2) \\ x_p &= f_x u_p + x_o \\ y_p &= f_y v_p + y_o \end{aligned}$$

where X, Y, Z are the 3D coordinates and x_p, y_p are the pixels position in the image plane. Figure 3 shows the image before (a) and after (b) correcting for radial distortion. We also compute the relative 3D position of one camera with respect to the others. This relative position remains the same and is useful for the construction of the mosaic.

Once we have calibrated CAMEO (intrinsic and extrinsic



Fig. 3. a) Original image b) Corrected image. Straight lines in the world are mapped to straight lines in the image.

parameters between cameras), we project the images into cylindrical coordinates to construct the mosaic [4][5]. Knowing the relative position between cameras, we can merge the images into the cylindrical plane. Figure 4 shows the original five images (a) and how they are merged into the mosaic (b).

To construct the mosaic, we make use of the highly optimized Intel Performance Primitive (IPP) libraries. The capture/mosaic process runs at 30 fps on a 3GHz PC running Microsoft Windows XP. This speed is achieved by constructing a look-up table (LUT) that maps each pixel in the mosaic to the position of one camera pixel and use a simple bilinear interpolation. The mapping takes into account the radial distortion, the cylindrical mapping and the relative position between cameras and it is computed once per camera. We are currently working on reducing parallax effects and automatically correct for small misalignments.

IV. FACE DETECTION

Automatic face detection is a well-known but difficult task due to the amount of variation in visual appearance of faces (faces vary in size, shape, coloring, and in small details such as the presence or absence of glasses, facial hair, hair style, etc). We use the face detector developed by Schneiderman and Kanade [6]. This algorithm is a parts-based method for classification of image regions into "face" and "non-face" regions. It explicitly models and estimates the posterior probability $P(\text{face}|\text{image})$ by choosing a functional form of the posterior probability function that models the joint statistics of local appearance and position on the face and the statistics of local appearance in the visual world. The algorithm uses approximately a million patterns to represent local appearance and counts the frequency of occurrence of these patterns over a large set of training images to compute the probabilities $P(\text{face}|\text{image})$. These probabilities are then used to classify an input window. This face detection method has high detection rates and low false positive rates but when run over the entire image, the face detection process is time consuming and hence not suitable for near real time applications. Examples of captured faces are shown in Figure 5.

In order to determine the location of people in the image, each subsequent frame of video is subtracted from the previous

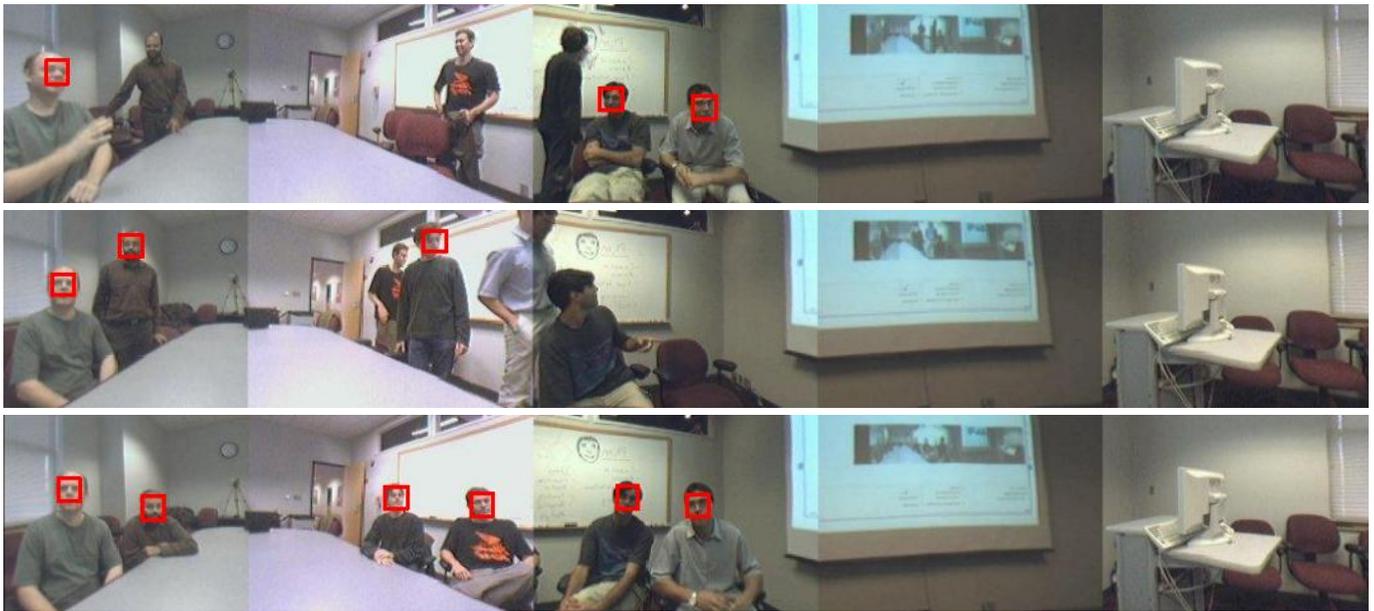


Fig. 5. CAMEO detects frontal faces with the Schneiderman and Kanade [6] face detection algorithm.

in order to determine what has changed. The pixels that represent motion are allowed to “persist” for several frames after they appear so that motion of objects in the image will be accentuated over time. Once the moving object regions are obtained, we perform some basic morphological operations such as erosion and dilation on the foreground image to reduce noise and then perform a connected component analysis on the foreground image to obtain motion blobs. We generate motion “regions of interest” (ROIs) by merging overlapping blobs and limit the search for faces to within the ROIs instead of over the entire image. Figure 6 illustrates the process of CAMEO’s face detection.

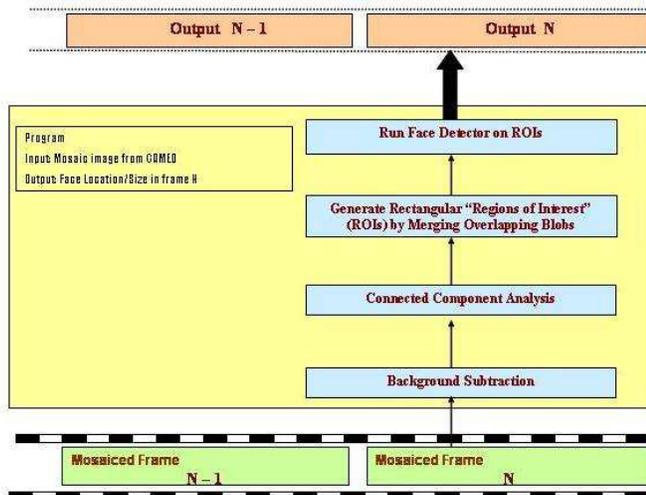


Fig. 6. Illustration of CAMEO’s face detection process.

V. LOGGING & GUI

CAMEO records each video data stream to an MPEG-4 movie file that can be used for off-line processing. CAMEO includes an on-line logging facility by which the tracked information is broadcast over the network to client for real-time monitoring, recording, and analysis. Meeting information can be recorded and used for many different purposes, including generating meeting summaries, allowing for specific queries about the meeting, and for possibly learning specific dynamic meeting patterns.

Several GUI tools, one of which is shown in Figure 7, have been developed for CAMEO which allow a user to augment the captured video stream with information such as who was attending the meeting, and when they arrived and if they left early. Because the current frontal face detection system only operates when people are looking directly at the camera, some additional processing is done to track people’s positions in the image. When CAMEO is told ahead of time how many people are attending the meeting, the correlation problem becomes much more tractable. Additionally, the GUI allows a user to label each of the people who attend the meeting if that information is not readily available.

CAMEO’s tracker system makes several basic assumptions about the dynamics of people in the meeting. One assumption is that during a meeting people typically do not move very fast as they are sitting down for the majority of the meeting duration. Even if someone is standing up and giving a presentation, they will remain in roughly the same place. Additionally, CAMEO uses the information about how many people are attending the meeting to constrain the matching (i.e. discarding the occasional false face detection).

Given these assumptions, a simple greedy heuristic is used to match the positions of people in the image from one frame

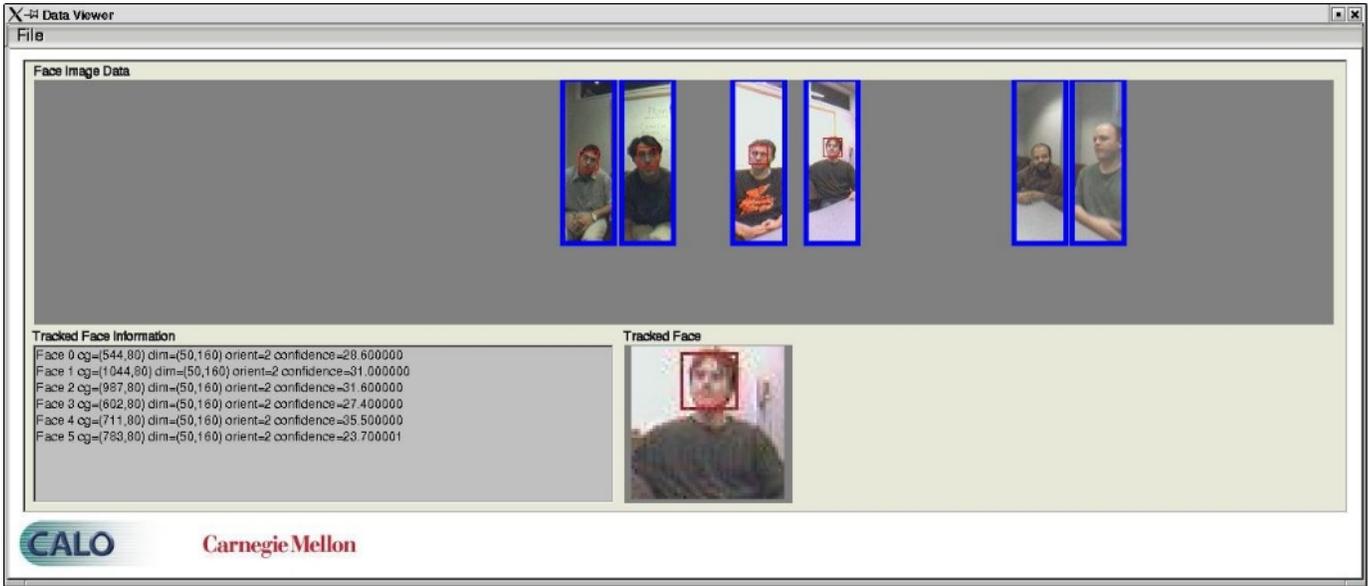


Fig. 7. CAMEO GUI illustrating the off-line replay and tracking system.

to the next, as shown in Figure 8. While in general, this tracker would typically lose track of people if they moved around a great deal, particularly if they occluded each other and moved very quickly. However, in practice, this first-order approximation works well in the meeting environments we have observed thus far. This is very useful because the computational overhead for this tracker is extremely low as compared to a system that makes use of face recognition to correlate the faces.

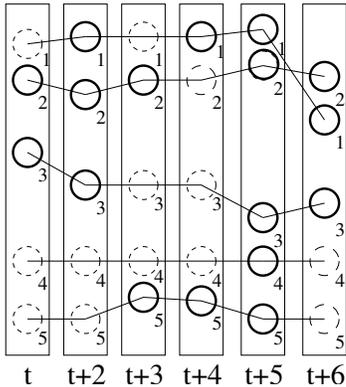


Fig. 8. A conceptual example of how the tracker correlates face positions from one frame to the next. Each column represents the faces found in an image at each timestep. If a face is not seen in an image (circles with dashed lines), its position is predicted by its position the previous frame. In the last frame, the relative positions of the faces change positions but are tracked because of the assumption that the most likely assignment is the motion that moves the least.

VI. RELATED WORK

Many techniques exist for generating coherent mosaics from several cameras [3][7][8][5][9]. However, most of them assume that the camera is panning or that there exists only

rotation between the camera's optical centers. In the case that the displacement between the optical center of the cameras is just rotational, it is easy to show that a homography can relate the geometry of the images taken from different cameras [10]. In our case, we must compensate for both a translation and a rotation between cameras.

Pfinder [11] is a real time system for tracking people which uses a multi-class statistical model of color and shape to obtain a 2D representation of the head and the hands. It models the background by observing the scene without people for a long time to estimate the color covariance associated with each pixel and then detects people by watching for deviations from this model. Another system that uses color and shape cues for tracking faces [12] is based on statistical color modeling and the deformable template. Our system does not require the storage of the background first. It also makes use of a face detection algorithm for identifying people which obviates the need for complex color histograms.

A real time visual surveillance system for detecting and tracking people that uses no color information but relies on a shape analysis and tracking to locate people and their parts is described in [13]. In our case, CAMEO will be used in a meeting situation where people's bodies are typically mostly occluded and so the face information is really the only body part that can be reliably detected and tracked.

The Video Surveillance and Monitoring work at Carnegie Mellon University [14] detects moving targets by frame differencing and tracks them by using a combination of temporal differencing and template matching. This work is most closely related to ours in the techniques that it employs, however, it primarily focuses on tracking people outdoors from a great distance while we are interested in tracking people indoors where people's bodies are occluded.

Body tracking work done at MIT [15] involve the use of

stereo cameras to determine the location of people's bodies and limbs. This information is then used to calculate where they are gesturing. We cannot make use of this technique due to the inherently monocular camera system that we are using.

VII. ONGOING/FUTURE WORK

In future work, CAMEO will be augmented to handle the case where the frontal face detector doesn't detect faces. In this case a simple model of an elliptical shape template combined with a human skin color model, shown in Figure 9, along with a motion blob detector (based on background differencing) will be used to detect the appearance of a person in the scene. This model will be used to generate a richer set of perceptual features that CAMEO will use to attempt to recognize and classify people's behaviors.

The head model consists of an elliptical template representing the top portion of the head (with the semi-major axis being b and semi-minor axis being a) combined with a horizontal line representing the shoulder (h being the vertical offset from the center of the ellipse and $2l$ being the width of the shoulder). Once a person's head is detected the body position will be estimated by the motion blob based on a simple face-body model.

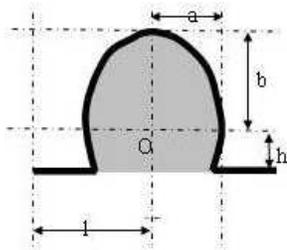


Fig. 9. Head model for the enhanced CAMEO tracking system.

The model color histogram of the head region and body region will be learned. These color models will be used to track the face and body in successive frames. Initial face/body estimates should be fully contained within the actual face/body area. For faces/bodies that are detected, the continuity of the person's motion will be exploited to limit the expected frame to frame motion. In particular, after each frame the velocity and acceleration of the motion models describing each person's position will be updated. We will use motion models to predict each person's location in the next frame. This predicted information will be fed to the tracker. We have implemented a tracker based on mean-shift analysis for mean-shift color tracking. The spatial gradient of this similarity measure is used to guide a fast search for the best candidate. This method is well suited for real time tracking applications. The system uses simple occlusion analysis to detect occlusion of one person by another and searches for the reappearance of the occluded person while maintaining track of the non-occluded persons. On reappearance, the system resumes track of the previously occluded person.

We have presented CAMEO, the Camera Assisted Meeting Event Observer, a physical awareness system designed to be

used by cognitive assistant systems to monitor actions during meeting settings. We have developed a high-resolution omnidirectional vision system that uses a face detection algorithm to locate the positions of people in an image. Higher level information about the meeting, such as the number of people in attendance, as well as a simple model of typical meeting behavior, is used to keep track of the positions of people through the meeting.

ACKNOWLEDGEMENTS

We would like to thank Takeo Kanade, Betsy Ricker, and Francesco Tamburrino, for their help with this project.

This research was supported by the National Business Center (NBC) of the Department of the Interior (DOI) under a subcontract from SRI International. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, by the NBC, DOI, SRI or the US Government.

REFERENCES

- [1] S. Baker and S. K. Nayar, "A theory of catadioptric image formation," in *Proceedings of the International Conference on Computer Vision*, 1998, pp. 35–42. [Online]. Available: citeseer.nj.nec.com/baker98theory.html
- [2] R. Benosman and S. B. Kang, Eds., *Panoramic Vision: Sensors, Theory and Applications*. New York: Springer-Verlag, 2001, ch. Single viewpoint catadioptric cameras.
- [3] P. Peer and F. Solina, "Panoramic depth imaging: Single standard camera approach," *International Journal of Computer Vision*, vol. 47, pp. 149–160, 2002.
- [4] R. Swaminathan and S. K. Nayar, "Nonmetric calibration of wide-angle lenses and polycameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1172–1178, 2000.
- [5] R. Szeliski and H. Shum, "Creating full view panoramic image mosaics and environment maps," *Computer Graphics*, vol. 31, no. Annual Conference Series, pp. 251–258, 1997.
- [6] H. Schneiderman and T. Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1998, pp. 45–51.
- [7] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *ACM Multimedia*, 2002.
- [8] J. Foote and D. Kimber, "Flycam: Practical panoramic video and automatic camera control," in *IEEE International Conference on Multimedia and Expo*, vol. 3, 2000, pp. 1419–1422.
- [9] S. Peleg and J. Herman, "Panoramic mosaics by manifold projection," in *IEEE CVPR Proceedings*, 1997.
- [10] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.
- [11] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997. [Online]. Available: citeseer.nj.nec.com/wren97pfinder.html
- [12] F. J. Huang and T. Chen, "Tracking of multiple faces for human-computer interfaces and virtual environments," in *IEEE International Conference on Multimedia and Expo*, New York, July 2000.
- [13] I. Haritaoglu, D. Harwood, and L. Davis, "Who, when, where, what: A real time system for detecting and tracking people," in *In Proceedings of the Third Face and Gesture Recognition Conference*, 1998, pp. 222–227.
- [14] A. Lipton, H. Fujiyoshi, and R. Patil, "Moving target classification and tracking from real-time video," in *Proc. of the IEEE Image Understanding Workshop*, 1998, pp. 129–136.
- [15] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 175–185, June 2000.