Proceedings of 2004 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 28 - October 2, 2004, Sendai, Japan

People Detection and Tracking in High Resolution Panoramic Video Mosaic

Raju Patil, Paul E. Rybski, Takeo Kanade, Manuela M. Veloso
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA, 15213
Email: {raju,prybski,tk,mmv}@cs.cmu.edu

Abstract— **We have designed a physical awareness system called CAMEO, the Camera Assisted Meeting Event Observer, which consists of a multi-camera omnidirectional vision system designed to be used in meeting environments. CAMEO is designed to monitor the activities of people in meetings so that it can generate a semantically-indexed summary of what occurred in the meeting. In this paper, we describe CAMEO's fast people detection and tracking module. This module makes use of a combination of frame differencing, face detection, and adaptive color blob tracking based on mean shift analysis to detect and track people in the panoramic image. We describe this algorithm and present experimental results from captured meeting logs.**

I. Introduction

Automatic detection and tracking of people in video is a challenging task with many applications such as surveillance and automatic video-indexing. This capability is a vital part of a robust physical awareness system designed to sense human beings in real world situations. We have designed a system called CAMEO, the Camera Assisted Meeting Event Observer, shown in Figure 1, which is an omnidirectional video system consisting of a set of several cameras. These cameras are oriented in such a way as to capture partially overlapping video frames. These frames are stitched together in real time to produce a high resolution panoramic mosaic. The system design and details of the mosaic generation are presented in [1].

In order to enable fast detection of people in the large mosaic image (1764×357 pixels), we employ frame differencing of subsequent frames of video to identify small subregions likely to contain moving people. We search these sub-regions for faces using a template-based face detection algorithm. Once located, we track the faces in the full image using a mean shift based color tracking algorithm. The person detection and tracking system then passes this information to a dynamic Bayesian network-based activity recognition system for the purposes of characterizing the motions of people in the environment and generating an indexed representation of the events that occurred within the meeting itself.

This paper is organized into the following sections. Related work is described in section II. The architecture of the detection and tracking system is described in section III. Results are presented in section IV. Conclusions and future work are presented in section V.

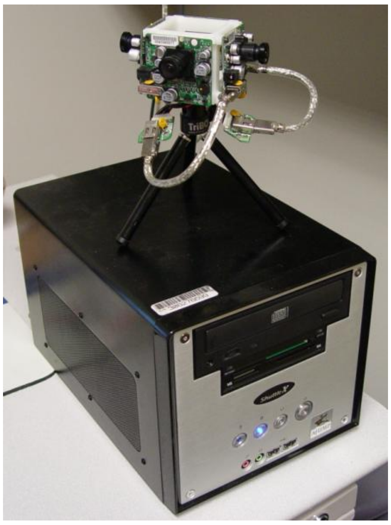


Fig. 1. The complete CAMEO system consists of several calibrated firewire cameras and a portable image-processing workstation.

II. Related Work

Various approaches have been tried for people tracking. Systems employing static cameras usually make use of background subtraction to detect and track people. A real time system called Pfinder is described in [2]. 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. It uses a multi-class statistical model of color and shape to model the head and the hands of people. Another system employing color and shape information for tracking faces [3] is based on statistical color modeling and the deformable template. Although our system makes use of static cameras in indoor environments, we cannot make use of background subtraction as CAMEO is a portable system designed to be used in different meeting rooms and typically, the system is started after the participants of the meeting are already inside the room, thus making the task of obtaining a background image very difficult. Our system does not require the storage of the background first and while we also use color based tracking like the above two papers, we employ mean shift analysis based on the work of [4] which enables faster tracking.

W4 [5] is a real-time visual surveillance system for de-

0-7803-8463-6/04/$20.00 ©2004 IEEE

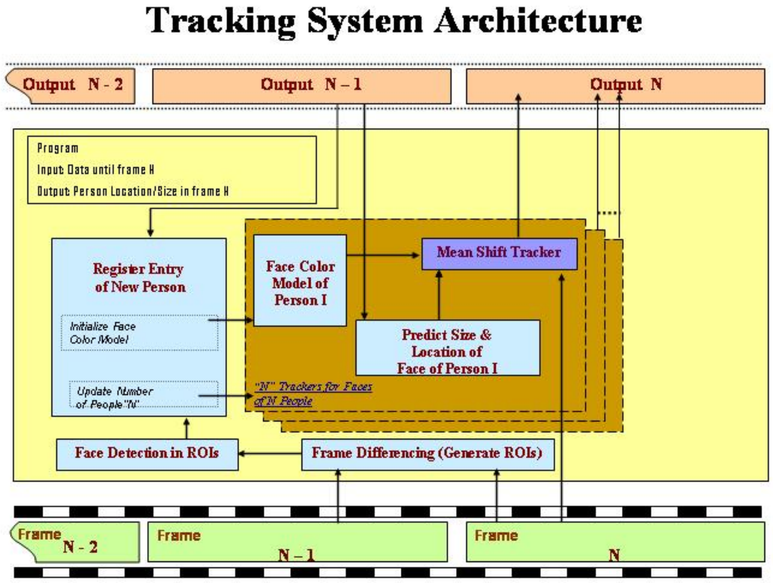

Tracking System Architecture
Output N - 2
Output N – 1
Output N
Program
Input: Data until frame N
Output: Person Location/Size in frame N
Register Entry of New Person
Initialize Face Color Model
Update Number of People "N"
Face Color Model of Person I
Mean Shift Tracker
Predict Size & Location of Face of Person I
"N" Trackers for Faces of N People
Face Detection in ROIs
Frame Differencing (Generate ROIs)
Frame N - 2
Frame N – 1
Frame N

Fig. 2. Diagram of the tracking system architecture. Regions of Interest (ROIs) are detected by subtracting subsequent frames of video. In typical indoor situations, usually motion and occasionally localized intensity changes generate ROIs which are passed to the face detector. If a face is detected, the color histogram is stored and used to track that face from that point on.

tecting and tracking people. They use no color information but rely on shape analysis and tracking to locate people and their parts. In our case, usually, only the torso and the face are visible in a meeting scenario and very often, the torso is also obscured by objects such as laptops. Currently, we only track people's faces instead of locating various body parts.

Previous work done at CMU for the Video Surveillance and Monitoring project [6] employs frame differencing for moving target detection and tracks them by using a combination of temporal differencing and template matching. We also employ temporal differencing but we use it to quickly detect areas of interest to limit the search area for the face detector, and we use color blob tracking instead of template matching.

Stereo vision is an important constituent of many methods such as [7]. In view of the current design of CAMEO, we do not use stereo. We largely rely on mean shift based color tracking and combine it with a few techniques to demonstrate successful detection and tracking.

Research in human/agent activity recognition is spread across a variety of different areas. On one side is gesture recognition, which attempts to use sensor input and signal processing techniques to recognize arm or hand gestures such as sign language [8]. On the other side is plan recognition [9] which ultimately attempts to classify a high-level set of goals, intentions, or belief states about agents (human or otherwise).

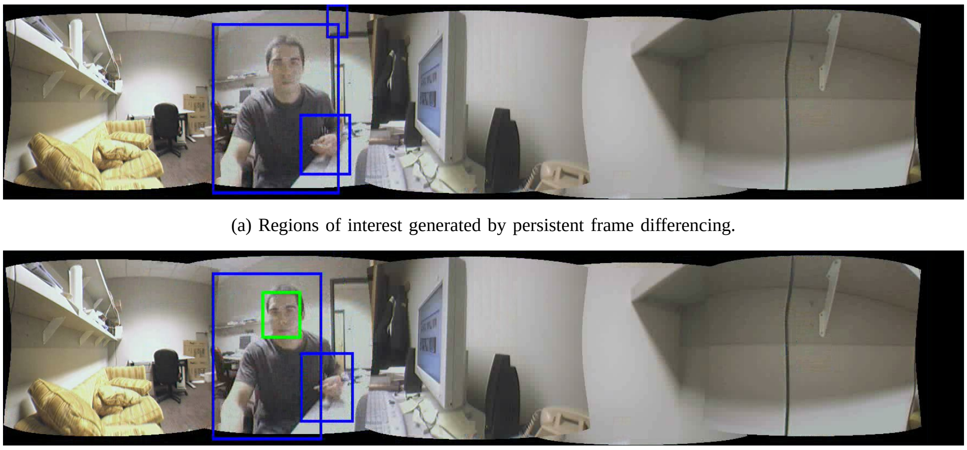
Dynamic Bayesian networks are used by [10] to recognize the gestures such as writing in different languages on a whiteboard, as well as activities such as using a Glucose monitor. However, instead of attempting to classify the specific kinds of actions that a human is doing, which tend to be very viewpoint dependent, we infer body stance and motion by tracking the user's face. This is a more general method of tracking and works well with the notion that CAMEO is designed to be set up and operated in relatively unstructured environments.

A system called the Abstract Hidden Markov mEmory Model (AHMEM) [11] is used to represent both state-dependent and context-free behaviors. This model represents a hierarchy of behavioral information ranging from lower-level sensory information up to a higher-level behavioral description. However, this work uses a network of cameras set up throughout the entire office space. Additionally, all of the locations in the workspace need to be labeled appropriately so that the system can reason about them.

III. Detection and Tracking System Architecture

There are several visual cues, such as motion, color, and face pattern, that are useful for person detection and tracking. Individually, these cues are not sufficient for tracking people in real life scenarios. Color and motion cues alone are not reliable enough and face pattern analysis over the entire image is computationally expensive. Therefore, we use all these cues in complementary ways to satisfy our real-time constraints. Figure 2 shows a schematic diagram of the vision system, indicating the interconnections between the various software components.

The following sections describe the details of the various components of CAMEO's tracking system. The components are the region of interest (ROI) extractor, face



(a) Regions of interest generated by persistent frame differencing.

(b) A face detected by running the face detector within the regions of interest.

Fig. 3. The process of detecting faces within the image. First, subsequent frames of video are subtracted from each other. The resulting pixels are grouped into regions of interest (ROIs) and passed to the face detection module.

detector, Ω shape detector, mean shift color tracker, and Bayesian network-based action recognizer.

A. Region of Interest Extractor

The CAMEO system generates large mosaic images (1764×357 pixels) which makes naively searching for faces across the entire image too time consuming. In order to enable fast detection of people in the large mosaic image, CAMEO employs persistent frame differencing based on the concept of "motion history images" developed in [12]. This frame differencing extracts small sub-regions likely to contain moving entities. The pixels of these sub-regions are allowed to "persist" for several frames after they appear so that the leading and trailing edges of moving objects are thickened over time. Once the thick edges of the moving objects are obtained, we perform some basic morphological operations such as dilation and erosion on the foreground image to fill holes and eliminate noise and then perform a connected component analysis on the foreground image to obtain blobs. We generate ROIs by merging overlapping blobs. Figure 3(a) shows the output of the ROI generator on one frame of the mosaic image.

B. Face Detection

Automatic face detection is extremely challenging due to the amount of variation in the size, shape, and color of faces. Additional difficulty is introduced by small details such as the presence or absence of glasses, facial hair, hair style, etc. We use the face detector developed by Schneiderman [13] which is a parts-based method for classification of image regions into ”face” and ”non-face” regions. It models and estimates the posterior probability $P(face|image)$ 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 a large set of training images to compute the probabilities $P(face|image)$. These probabilities are then used to classify an input window. This face detection method achieves high detection rates and low false positive rates. Figure 3(b) shows the face discovered in the ROIs caused by the motion of the person in the image.

C. Mean Shift Based Color Tracking

Once faces are detected, we track the location of the faces in subsequent frames using mean shift based color tracking. Color is a very useful cue for tracking non-rigid entities such as people in video sequences. Color based tracking is robust to rotation in depth, partial occlusion and clutter. We have implemented a method based on work presented in [4]. A color histogram of the face can be learned in the frame in which a person's face has been detected and the spatial position of this histogram can be tracked in successive frames. The Bhattacharya coefficient [4] is used as the similarity measure between the model color histogram and the target color histogram. The spatial gradient of this similarity measure is used to guide a fast search for the best candidate. The optimization, based on mean-shift analysis, converges in only a few iterations and is thus well suited for real-time tracking. A 1D histogram of a desired part of an input image can be computed by considering all the possible colors formed by the quantized RGB values. We choose 4-bit quantization of the RGB values. This gives us $n = 2^4 \times 3 = 4096$ bins. Given such a n-bucket model histogram $m_i | i = 1 \ldots n$ (learned from the first frame after detecting a person) and a data histogram $d_i | i = 1 \ldots n$, computed from a candidate location in a successive frame, we compute a similarity measure using the Bhattacharya coefficient, ρ, as follows:

$$\rho = \sum_{i=1}^{n} (\sqrt{m_i d_i}) \qquad (1)$$

As pointed out in [4], this similarity measure has the following desirable properties; (a) it is nearly optimal, (b) it imposes a metric structure, (c) it is scale-invariant, i.e., invariant to object size (number of pixels), and (d) it is valid for arbitrary distributions (not just Gaussian). This measure

ρ is maximized by using standard mean-shift iterations to obtain the new location in the current frame which best matches the model histogram. In order to handle scale change in the video sequence, we change the window size by $\pm 10\%$ and choose the size which yields the best similarity measure.

For faces that are detected, we take advantage of the continuity of the person's motion; that is, the frame to frame motion will be limited. In particular, after each frame we update motion models describing each person's position and velocity. We then use these motion models to predict each person's location in the next frame and we feed this predicted information to our tracker.

The system uses simple occlusion analysis based on relative scores of the Bhattacharya coefficient on the body regions 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.

D. Ω Shape Detector

We use an additional shape template plus color based detector called the Ω detector to help reduce the number of false positives generated by the face detector. We check each detected face with our Ω detector and keep only those detected faces which have obtained a high score on the Ω detector test. The Ω template is based on the idea of an elliptical shape tracker presented in [14] which combines the shape cue (elliptical template) with color cues to model the face. We model the 2D projection of the human head and shoulder onto the image plane with an Ω shaped template. Figure 4 shows the Ω template with its various parameters. We also combine the Ω shaped template with a skin color classification module to model the human head and shoulder. These two cues, based on shape and color, are complementary in nature. In cases of partial occlusion of the head boundary causing loss of information for the shape module, the color module is successful whereas the shape module is successful when enough face color information is not available due to rotation or partial occlusion of the interior area of the face.

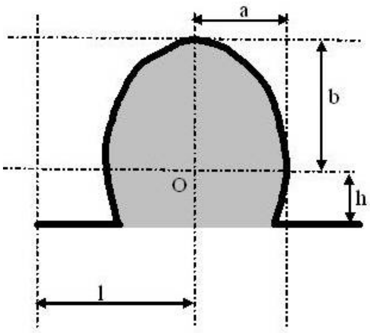


Fig. 4. Ω shaped template for 2D projection of head and shoulder

We use an elliptical template underlying our Ω shaped template, with the major axis parallel to the vertical axis and the minor axis parallel to the horizontal axis. As in [14], we choose a fixed aspect ratio of 1.2 (i.e., $b = 1.2(a)$). The other two parameters of the ellipse are the offset h of the shoulder from the center of the head (O) and the horizontal extent of the shoulder, l. In the area of the image corresponding to the faces detected by the frontal face detector, we search for the best matching Ω template based on two scores; (a) the gradient matching score which measures how well the boundary of the candidate image patch agrees with the Ω shape and (b) the color matching score which measures what proportion of the shaded area (see Figure 4) consists of skin color pixels. We retain only those detected faces which return a high combined Ω detector score.

E. Dynamic Bayesian Network Action Classifier

Because CAMEO is designed for use in unstructured environments where nothing is known about the positions of people in the room, we make the assumption that the only feature CAMEO can reliably track is the head and face. Due to uncertainty in depth and occlusions, we attempt to infer high-level activities solely in terms of head motion since this is a fairly good indicator of body position. Instead of attempting to solve the image understanding problem purely from data, we construct a set of dynamic Bayesian network classifiers from *a priori* knowledge about meetings and the interactions between people in those meetings. In this system, we attempt to classify simple states such as “sit”, “stand”, “sitting down”, and “standing up”. The allowable transitions in this network are defined by a simple finite state machine which is encoded into the conditional probability state transition table (CPT) of a dynamic Bayesian network.

Dynamic Bayesian networks are directed acyclic graphs (DAGs) that model stochastic time series processes. They are a generalization of both Hidden Markov Models (HMM) [15] and linear dynamical systems such as Kalman Filters. DBNs represent both the hidden and observed system state in terms of state variables whose representations are described as part of the DBN's node topology. DBNs are defined by an initial state distribution, a state transition model, and an observation model. Combining the prior model with the transition model creates what is called a two-slice temporal Bayes net (or a 2TBN) [16]. Figure 5 illustrates a simple DBN representation of a HMM. The hidden states are the X nodes while the observations are the Y nodes. The model is unrolled over time so that each time “slice” of the model can be observed as a distinct set of nodes.

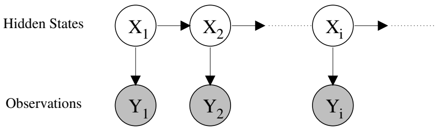


Fig. 5. Simple Dynamic Bayesian Network (HMM) with a single hidden node (the model is “unrolled” over time).

We wish to determine the value of hidden state from the observation at each timestep. Because the hidden states in the networks used in this paper are discrete, the inference procedure is identical to the inference method for HMMs called the forward-backwards algorithm [15].

IV. Results

The system was tested on short video sequences containing a few people moving around in a meeting environment. Figure 6 shows a few frames of video captured from one such sequence. Once detected and tracked, each person's face is labeled with a letter (A-C). In this video, the three people's faces move around the image and change direction from frontal to profile. The tracker successfully monitors their positions and maintains the appropriate label. While there are no occlusions in this video, it is fairly indicative of a typical meeting environment in which most of the participants are in view of CAMEO. Performance-wise, our system detects and tracks people in CAMEO's high resolution mosaic video frames ((1764×357 pixels) at 3 frames per second on a 3 GHz Pentium 4 machine.

Figure 7 shows the results from the dynamic Bayesian network action recognition system. In this figure, the solid line shows the hand-labeled ground truth of the person's activities, the dashed line shows the estimated activities, and the circles indicate states that were misclassified. Of the 315 images encoded with person tracked data, only 29 of the states were misclassified. Most of these occurred during the transitions from the "stand" state through the "sitting down" state to the "sit" state. This is primarily due to variances in the way that people move around. Incorporating a larger collection of example person states would help to alleviate some of these misclassifications.

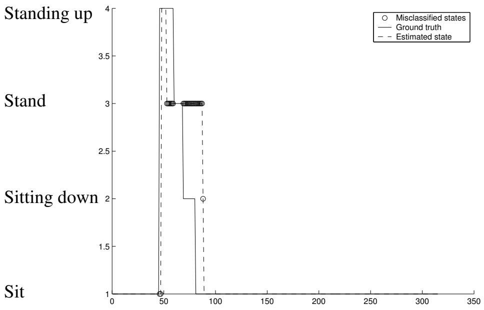


Fig. 7. Classified actions from the center person in the dataset illustrated in Figure 6. Of the 315 timesteps in this dataset, only 29 of the states were misclassified. The Y-axis represents the state where, 1="Sit", 2="Sitting down", 3="Stand", and 4="Standing up". The X-axis is the frame count. The data was captured at 15 fps.

V. Conclusion and Future Work

We have developed a system for detecting and tracking people that combines various techniques (frame differencing, pattern matching, and color based tracking) to achieve fast person tracking in panoramic mosaiced video of typical indoor scenarios. Our approach focuses on a technique that works in completely unstructured environments where we assume that we will only be able to see people's faces and heads. In future work, we will be including the tracking of additional body parts such as the torso and limbs in our system and we are working on adding other vision cues such as optic flow and template matching to achieve more robust tracking performance. Additionally, we are implementing a multi-hypothesis tracking system that will combine the color histogram information with simple motion models of the people in the video sequence to make the system more robust to occlusions and false positives. Adding this kind of tracking system will also help resolve ambiguities caused by people occluding each other as they move around the environment. Finally, we are using CAMEO to collect a large corpus of meeting data that will be used to train the action-recognition classifier such that it will recognize a larger set of actions and handle wider person-specific variations of those actions.

Acknowledgment

We would like to thank Carlos Vallespi, Fernando De La Torre, Daniel Hershey, Jonathan A. White, Brett Browning, 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] P. E. Rybski, F. de la Torre, R. Patil, C. Vallespi, M. Veloso, and B. Browning, "Cameo: Camera assisted meeting event observer," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Apr 2004.

[2] 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

[3] 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.

[4] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2000, pp. 142–151. [Online]. Available: citeseer.nj.nec.com/comaniciu00realtime.html

[5] 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.

[6] 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.

[7] 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.

[8] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," in *International Symposium on Computer Vision*, vol. 5B Systems and Applications, November 1995, pp. 265–270.

[9] J. Allen, H. Kautz, R. Pelavin, and J. Tennenberg, "A formal theory of plan recognition and its implementation," in *Reasoning About Plans*. Morgan Kaufmann Publishers, 1991, ch. 2, pp. 69–126.

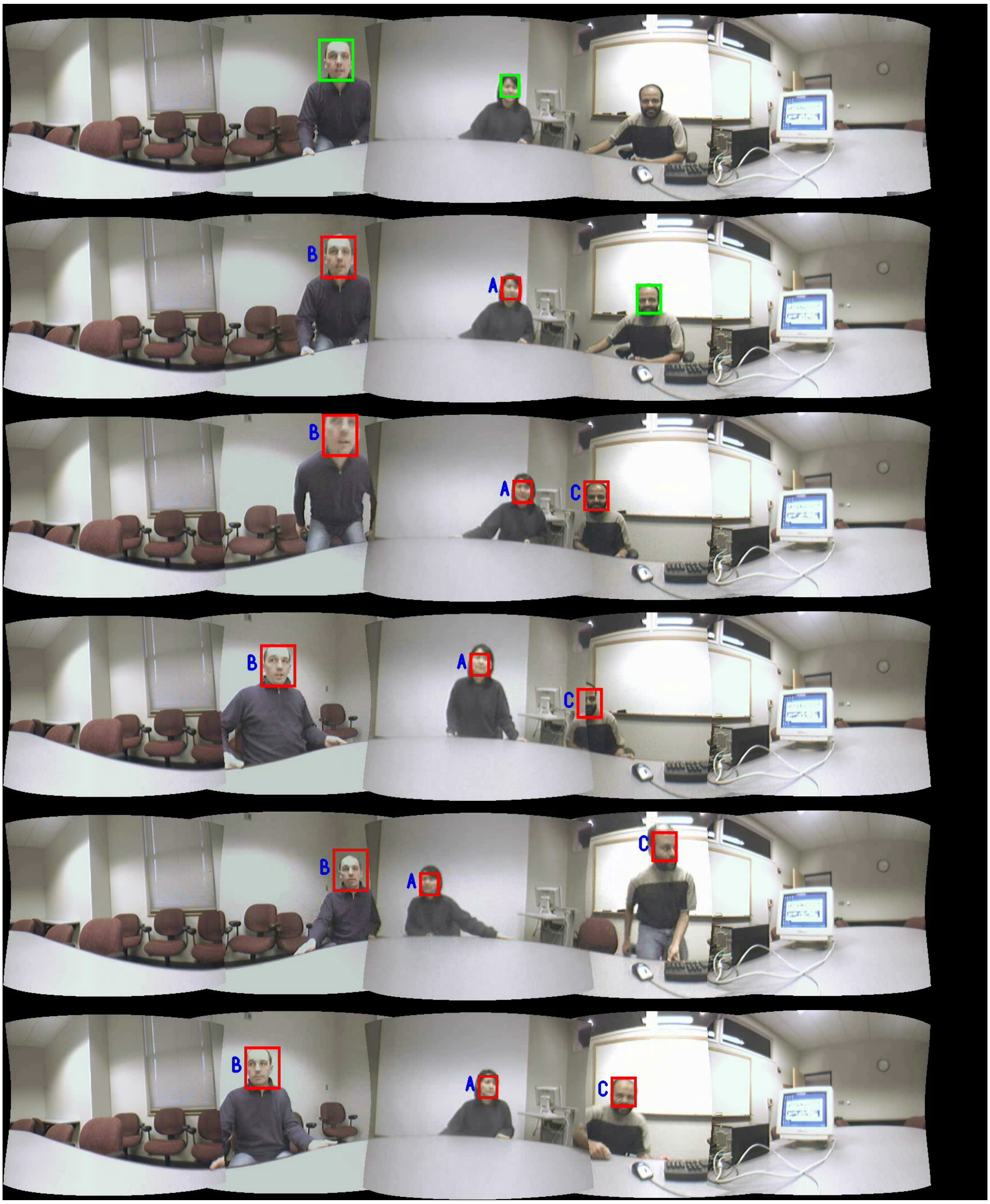


Fig. 6. CAMEO's face detection and tracking results. Frames 1,2,20,60,150, and 200 are shown above. Persons "A" and "B" were detected in frame 1 and person "C" is detected in frame 2. All three persons are subsequently tracked through the sequence.

[10] R. Hamid, Y. Huang, and I. Essa, "ARGMode – activity recognition using graphical models," in *Conference on Computer Vision and Pattern Recognition Workshop*, vol. 4, Madison, WI, June 2003, pp. 38–44.

[11] N. Nguyen, H. Bui, S. Venkatesh, and G. West, "Recognizing and monitoring high level behaviours in complex spatial environments," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.

[12] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 1998, pp. 928–934.

[13] H. Schneiderman, "Feature-centric evaluation for cascaded object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[14] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," 1998. [Online]. Available: citeseer.nj.nec.com/birchfield98elliptical.html

[15] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[16] K. Murphy, "Dynamic bayesian networks: representation, inference and learning," Ph.D. dissertation, UC Berkeley, Computer Science Division, July 2002.