CAMEO : Modeling Human Activity in Formal Meeting Situations

Paul E. Rybski, Fernando de la Torre, Raju Patil, Carlos Vallespi, Manuela Veloso, Brett Browning
Robotics Institute,Carnegie Mellon University
5000 Forbes Ave.,Pittsburgh, PA 15213
Email: {prybski,ftorre,raju,cvalles,veloso,brettb}@cs.cmu.edu

Abstract

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 used to observe formal meeting environments and infer the activities of people attending them.

Introduction

We have recently engaged on the challenging development of an agent to assist users in everyday office-related tasks. In particular, the agent needs to keep track of the state of their users so it can anticipate the user's needs and proactively address them. To address this challenge, we have developed CAMEO, the Camera Assisted Meeting Event Observer (Rybski *et al.* 2004), which is a physical awareness system designed for use by an agent-based electronic assistant. CAMEO addresses the problem of extracting and reasoning about high-level features from real-time and continuous observation of a meeting environment. Contextual information about meetings and the interactions that take place within them are used to define Dynamic Bayesian Network classifiers to effectively infer the state of the users as well as a higher-level state of the meeting. CAMEO is part of a larger effort called CALO (Cognitive Agent that Learns and Organizes) to develop an enduring personalized cognitive assistant that is capable of helping humans handle the many daily business/personal activities in which they engage. CAMEO is intended to be used in a manner similar to a speaker/video phone for a conference call.

The CAMEO System

CAMEO is an omni-directional camera system consisting of four or five firewire cameras (CAMEO supports both configurations) mounted in a circle, as shown in Figure 1. The individual data streams coming from each of the cameras are merged into a single panoramic image of the world. The cameras are connected to a Small Form-Factor 3.0GHz Pentium 4 PC that captures the video data and does the image processing.

Faces are detected using an algorithm by Schneiderman and Kanade (Schneiderman & Kanade 1998) which is a parts-based method for classification of image regions into ”face” and ”non-face” regions. It explicitly models and estimates the joint statistics of local appearance and position on the face and the statistics of local appearance in the visual world. by using 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. Once detected, the faces are classified by computing a distance metric between sets of tracked faces and the new faces. Matches are those that have the smallest distance. The metric is computed by taking the SVD of the image sets and computing the weighted sum of the most significant eigenvectors. The relative displacements of the face's centroid are used as features for the person action recognition system. Finally, a color histogram of the person's face and torso is learned to allow the person to be tracked from frame to frame regardless of whether their face is visible to CAMEO.

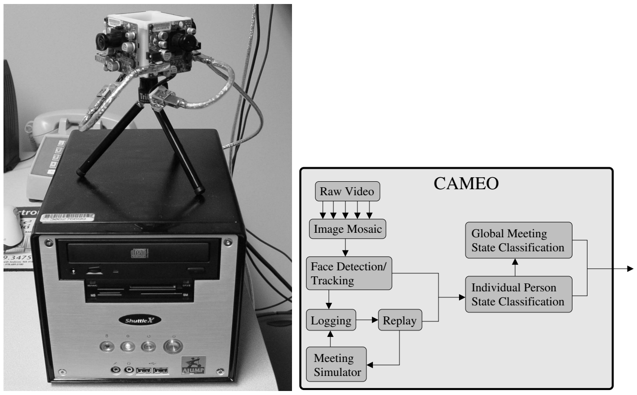
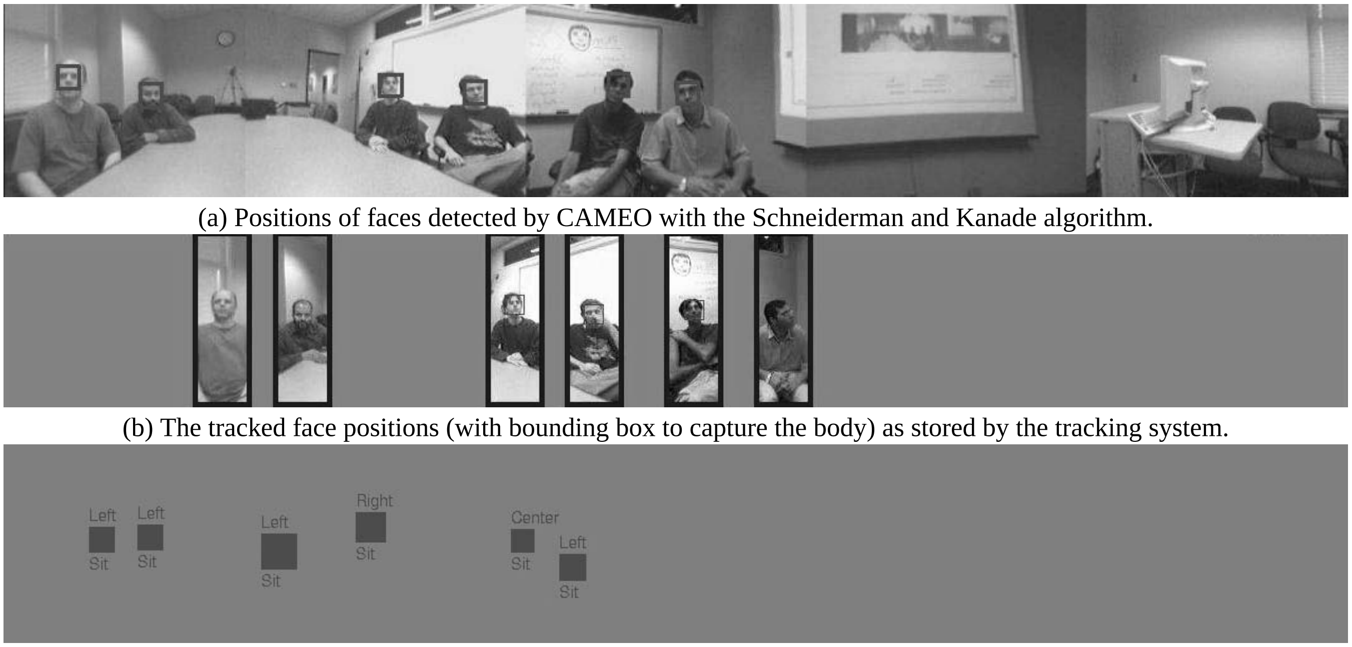


Figure 1: CAMEO consists a panoramic vision system and a small-form-factor PC. CAMEO processes video from its camera to locate people and classify their actions.

The panoramic video stream is scanned for human activity by identifying the positions of human faces found in the image. This low-level visual information is fed into a Dynamic Bayesian Network (Murphy 2002) (DBN) classifier system. The classifier determines the state of the individual people in the meeting. These individual person state esti-

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Positions of faces detected by CAMEO with the Schneiderman and Kanade algorithm.

(b) The tracked face positions (with bounding box to capture the body) as stored by the tracking system.

(c) Synthesized facial data from a simulated meeting whose parameters were learned from real data.

Figure 2: Several data gathering and representation layers in CAMEO.

mates are then used to infer high-level state estimates of the meeting itself. Our approach makes use of a very specific set of contextual information regarding the meeting domain to generate the Bayesian classification system, rather than attempting to solve the general image understanding problem.

Inferring Human Activity

Inferring the state of activities in a meeting takes place at two levels. The first level is the state classification of the individual people attending the meeting. The second level is classification of global meeting state, which is done after the individual states of the people are determined. 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 how interactions between people in those meetings.

In order to determine the state of each person in the meeting, a Dynamic Bayesian Network (DBN) model of a Hidden Markov Model (HMM) (Rabiner 1989) is created with a single discrete hidden node and a single continuous-valued observation node. Given a sequence of real-valued state observations from the meeting, the above DBN inference algorithm is used to infer the state of each person from data.

The state transition function for the DBN's hidden state is defined by a simple finite state machine model of a person's behavior. Example states for an individual person's actions include: stand, standing , sit, sitting, and walking. The conditional probability distribution for the hidden node are either hand-coded or learned from collecting statistics from CAMEO's observations. The global state of the meeting is determined by examining all of the states of the individual meeting participants. Allowable state transitions are defined as a first-order (fully-observable) Markov model which takes into account a minimum duration for a state transition.

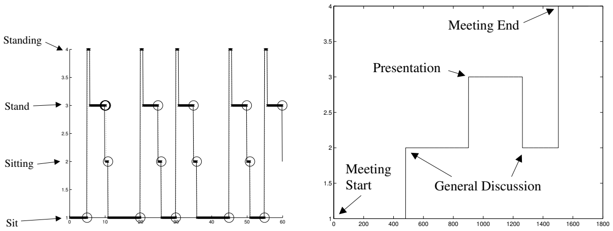


Figure 3: Classified states from a single person (left) as well as the global state of the meeting (right).

References

Murphy, K. 2002. *Dynamic Bayesian Networks: representation, Inference and Learning*. Ph.D. Dissertation, UC Berkeley, Computer Science Division.

Rabiner, L. R. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Rybski, P. E.; de la Torre, F.; Patil, R.; Vallespi, C.; Veloso, M. M.; and Browning, B. 2004. Cameo: The camera assisted meeting event observer. In *International Conference on Robotics and Automation*.

Schneiderman, H., and Kanade, T. 1998. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 45–51.