Inferring Human Interactions From Sparse Visual Data

Paul E. Rybski, Manuela M. Veloso
Robotics Institute,Carnegie Mellon University
5000 Forbes Ave.,Pittsburgh, PA 15213
{prybski,mmv}@cs.cmu.edu

Abstract

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. The state of the user may be easily available when the user directly interacts with their agent through a PC or PDA interface. However, when the user attends a meeting and interacts with other people, PC and PDA interfaces are not sufficient to give the agents a general view of the environment in which their users are interacting. In this paper, we introduce the CAMEO, the Camera Assisted Meeting Event Observer, which is a physical awareness system designed for use by an agent-based electronic assistant. We then present a particular aspect of CAMEO and main contribution of the paper, namely how 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 with them is 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. We present and show results of the state inference algorithm.

1. Introduction

Electronic agents designed to interact with humans and help them carry out their day-to-day business in an office domain require a good estimate of their user's state. Such a state estimate might consist of the projects the user is responsible for completing, the set of resources that the user has at his/her disposal, and the user's daily schedule, to name a few. By obtaining a good state estimate, an electronic agent will be able to reason about its user's needs and address them as best as it can. Ideally, it would be able to anticipate future needs and prepare for them.

Obtaining an accurate estimate of the user's state is a difficult challenge. Electronic agents that interact directly with humans (as opposed to those that might only handle email or scheduling information) can obtain information from a range of different sources including, traditional workstation/PDA input devices, spoken audio processing, and video processing systems. Workstation/PDA interfaces require that the user be using the device in question so that the data or queries/requests can be entered directly. Audio and video processing systems are more flexible in that the user can interact with an agent in a manner that is closer to interacting with a co-worker than with a data-entry device.

However, regardless of the interface, many important human interactions take place outside of the office and typically not in a fashion in which the agent can observe or participate. Of particular interest are interactions that take place in formal meeting environments. Having a sensor suite present in a meeting environment would give an agent useful information about what tasks it could assist its user with. Afterwards, this information could be automatically organized such that the agent could easily answer questions posed by its user 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?" An agent that could recognize these events would provide its user with the ability to recall events throughout the working day whose importance might have been initially missed.

In order to address this challenge, our group is developing a physical awareness system for an agent-based electronic assistant called CAMEO (Camera Assisted Meeting Event Observer) [12]. 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.

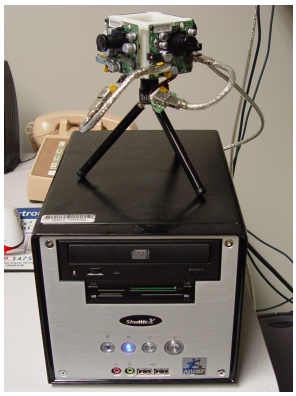


Figure 1. The CAMEO† system consists of a set of firewire cameras arranged in a panoramic fashion and a small-form-factor PC.

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 (DBN) classifier system. The classifier determines the state of the individual people in the meeting. These individual person state estimates 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.

2. Related Work

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 [14]. On the other side is plan recognition [1] which ultimately attempts to classify a high-level set of goals, intentions, or belief states about agents (human or otherwise). Our work falls somewhere between those two areas of research in that we are interested in inferring high level behavioral interactions (as restricted to a meeting domain) from fairly sparse sensor information.

Dynamic Bayesian networks are used by [5] to recognize the gestures such as writing in different languages on a whiteboard, as well as activities such as using a Glucose monitor. The gesture recognition system described in this work is probably the most similar to ours. 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 operate in relatively unstructured environments.

In [7], finite state machine models of gestures are constructed that by learning the spatial and temporal information of the gestures separately from each other. This allows for relatively complex spatial patterns (such as figure-8s) to be learned from the data. However, it is assumed that the gestures are performed directly in front of the camera and that the individual features of the face and hands can be recognized and observed without error.

An extension to the Hidden Markov Model [11] formalism called the Abstract Hidden Markov mEmory Model (AHMEM) [10] 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.

A system for torso and arm position tracking is described in [4]. This research makes use of stereo cameras to fit geometric models onto the torso and arms of a human so that they can communicate deictic information to the tracking system. Our system is essentially monocular and is not intended to be addressed directly where it could observe the full torso and arm positions of everyone attending the meeting.

Recognizing the behaviors of individual robotic (non-human) agents has been studied in [6]. Robots playing soccer against other robots [3] would greatly benefit by being able to classify the different behavioral patterns observed by their opponents. In this work, robots are tracked by an overhead camera and their actions are classified by a series of hand-crafted modified hidden Markov models (called Behavior HMMs). Each model has an additional accept state as well as multiple reject states. The relative displacements of the robots are observed over time and the model which fits the displacement sequence the best is chosen as the correct one.

In contrast to the previous approach, research has also been done in automatically extracting behavior sequence information from agent data and classifying it in a non-supervised fashion [8]. While this approach does not need *a priori* models such as in the previous approach, it does require the presence of semantically-labeled data to operate properly. This work was done primarily in a software agent domain where such information is more readily available.

† Special thanks to Fernando de la Torre, Raju Patil, Carlos Vallespi, Brett Browning and Betsy Ricker for their help with the development of CAMEO.

3. CAMEO as Part of an Agent-Based Office Assistant

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. This larger effort is engaged in developing a personalized omnipresent computation resource that can handle routine tasks and events, anticipate predictable user needs and prepare for them, and assist in handling unexpected events. CAMEO is intended to be used similarly to a speaker or video phone for a conference call. It is placed in the center of the meeting room where it has a relatively unobstructed view of the meeting participants. This is more flexible (and inexpensive) than instrumenting the room with stationary cameras. 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.

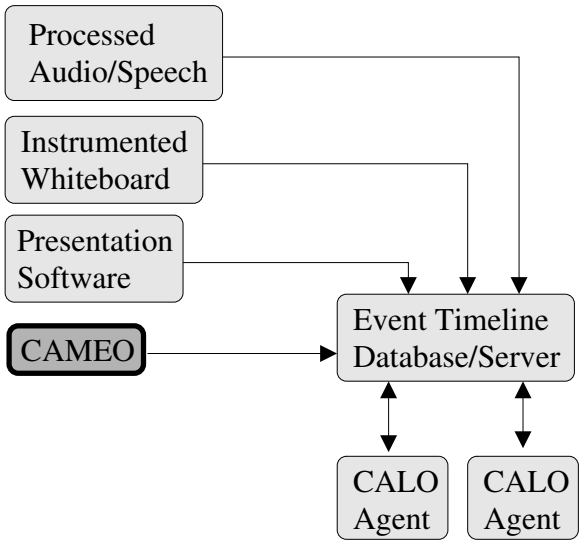


Figure 2. CAMEO captures live video from the meetings, determines what people are doing and posts this information to a centralized server. Other sources of meeting information server (not part of CAMEO) include audio processing, digitized whiteboards, and information from presentation software. Individual CALO agents query the server to access this information.

In order to be useful in a general set of environments, CAMEO must be able to operate in many sorts of meeting room configurations and should not require any lengthy calibration for distance or lighting conditions if possible. CAMEO must operate in uncalibrated environments where its position in the room and the positions of the meeting participants are initially unknown. Because CAMEO will not know ahead of time where people are sitting, or how much of the person will be visible at any given time, no complex body models are used for tracking purposes. The only assumption that CAMEO makes about people attending the meeting is that their faces will be visible.

Figure 2 illustrates how CAMEO operates in a CALO-enabled office environment. At the highest level, CAMEO is a data source provider which converts raw visual information captured in the meeting environment into semantically-labeled events that are stored in a centralized event timeline database/server. Other similar data sources (beyond the scope of this paper) could include system such as instrumented whiteboards that digitize what's written on them, as well as audio/speech processing systems to record and understand what is said. CALO agents belonging to the individual users connect to this timeline server and query it for the specific events of interest.

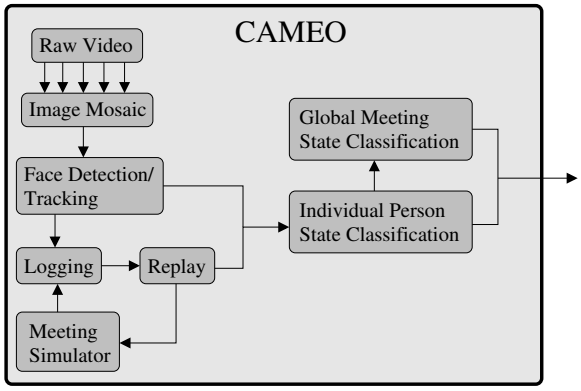
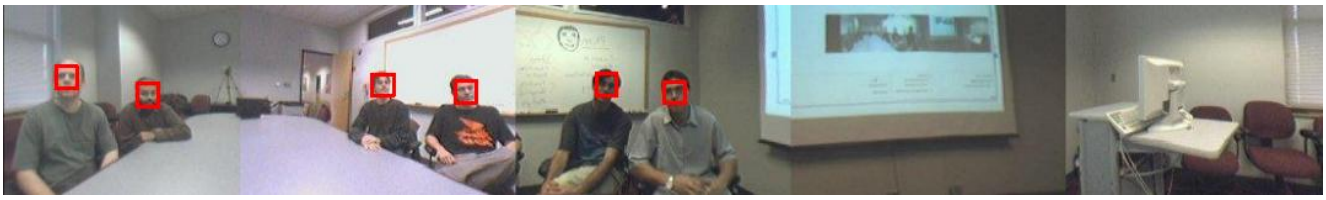


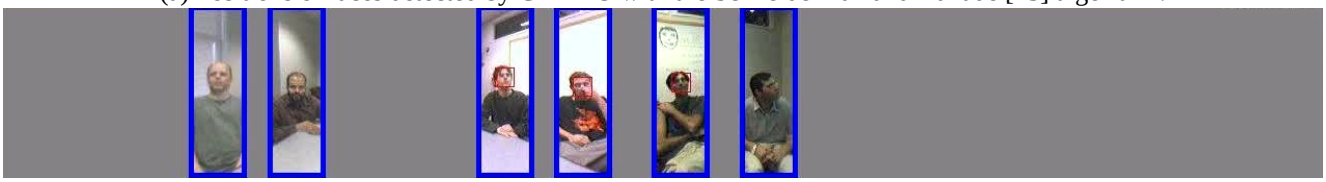
Figure 3. Interconnections between CAMEO's internal modules.

Interactions between CAMEO's internal modules is illustrated in Figure 3. Raw visual data from the multi-camera system is captured and merged into a single consistent image mosaic. People in the image are located by identifying their faces. CAMEO makes use of a face detector developed by Schneiderman and Kanade [13] which is a parts-based classifier that uses a database of over a million patterns to classify the presence of faces in the image. Examples of captured faces are shown in the Figure 4(a). Each person's face is located and tracked over time.

Faces are matched between subsequent frames by computing a distance metric between sets of tracked faces and



(a) Positions of faces detected by CAMEO with the Schneiderman and Kanade [13] algorithm.



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

(c) Synthesized facial data from the meeting simulator.

Figure 4. Several data gathering layers in CAMEO.

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.

The tracked faces are registered, as shown in Figure 4(b), and stored in a database. If CAMEO is used in a meeting where it does not have access to a network, the face information is logged for playback and analysis later. Once CAMEO is back on the network, the event information can be offloaded to the timeline server. Statistics from the stored face information are used to generate models for a meeting simulator, shown in the Figure 4(c), that can be used to synthesize the same kind of data that is captured from live video. This is useful for debugging and training the person and meeting state classifier system.

Streams of tracked facial information (from live video, archived data, or synthesized by the simulator) are fed into a Dynamic Bayesian Network classifier that classifies the state of each person. Specific events such as when a person changes state, such as when they stand up or sit down, are noted and passed on to a timeline server. The individual person state values are passed into a higher-level classifier which computes the global meeting state from the interactions of the group as a whole. This classifier also passes information to the timeline server so that global events such as the transition from a presentation to a general discussion are noted. The follow sections describe the details of CAMEO's state inference mechanisms.

4. Meeting State Inference

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 the classification of the global state of the meeting, 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.

4.1. Dynamic Bayesian Networks

For the sake of completeness, we will provide a short review of inference on time-series data using Dynamic Bayesian Networks (DBNs). DBNs are directed acyclic graphs (DAGs) that model stochastic time series processes. They are a generalization of both Hidden Markov Models (HMM) [11] and linear dynamical systems such as Kalman Filters. Similarly to these two models, 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. Following the formalism defined by Murphy [9], DBNs are defined by an initial state distribution $P(Z_1)$, a state transition model $P(Z_t|Z_{t-1})$, and an ob-

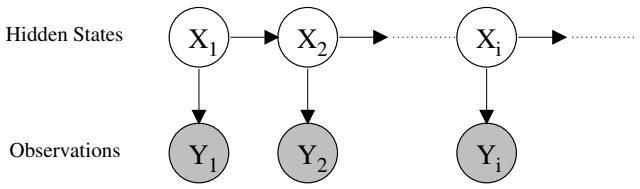


Figure 5. Simple Dynamic Bayesian Network in the form of a Hidden Markov Model with a single hidden node (the model is "unrolled" over time).

servation model $P(Y_t|Z_t)$, over all time $t = 1 \ldots T$, where Z_i is a random variable. The state transition model is defined as:

$$P(Z_t|Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^i|Pa(Z(_t^i))) \quad (1)$$

where Z_t^i is the *i*'th node at time t, N is the number of hidden states in each slice of the model, and $Pa(Z_t^i)$ represents the parents of variable Z_t^i in the graph (where we make the assumption that the parents of a node follow the first-order Markov assumption). Combining the prior model with the transition model creates what is called a two-slice temporal Bayes net (or a 2TBN). 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.

We wish to determine the value of hidden state from the observation at each timestep t, or to compute the value $P(X_t|y_{1:T})$. 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 [11]. This involves recursively computing a forward term $\alpha_t(i) \equiv P(X_t = i|y_{1:t})$ and a backward term $\beta_t(i) \equiv P(y_{t+1:T}|X_t = i)$ and combining them to produce $P(X_t = i|y_{1:T})$ via the formula:

$$P(X_t = i|y_{1:T}) = \frac{P(y_{t+1:T}|X_t = i)P(X_t = i|y_{1:t})}{P(y_{1:T})} \quad (2)$$

4.2. Person State Inference

In order to determine the state of each person in the meeting, a DBN model of a HMM 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 the data.

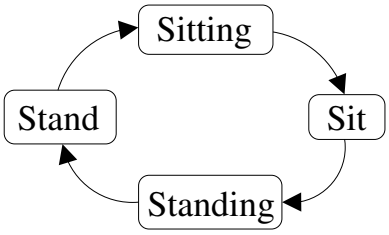


Figure 6. Example finite state machine for a single person in a meeting.

The state transition function $P(X_t|P_{t-1})$ for the DBN's hidden state is defined by a simple finite state machine (FSM) model of a person's behavior. As an example, Figure 6 illustrates a FSM that captures the set of possible states that a person could be in at any given time during a meeting. The "standing" state represents the action of changing from a sit position to a stand position, while the "sitting" state represents the action of changing from a stand position to a sit position. Additional states such as "walking" could be added to this machine, which would represent the action of moving from one point in the environment to another point. The possible values that the hidden state could take represent such values as "sit", "sitting", "stand", and "standing". The conditional probability distribution for the hidden node are either hand-coded or learned from collecting statistics from CAMEO's observations, as illustrated in the following table:

Conditional Probability	Learned Value
$P(X_t = \text{sit}\|X_{t-1} = \text{sit})$	0.99
$P(X_t = \text{standing}\|X_{t-1} = \text{sit})$	0.01
$P(X_t = \text{sitting}\|X_{t-1} = \text{sitting})$	0.87
$P(X_t = \text{sit}\|X_{t-1} = \text{sitting})$	0.12
$P(X_t = \text{stand}\|X_{t-1} = \text{stand})$	0.99
$P(X_t = \text{sitting}\|X_{t-1} = \text{stand})$	0.01
$P(X_t = \text{standing}\|X_{t-1} = \text{standing})$	0.85
$P(X_t = \text{stand}\|X_{t-1} = \text{standing})$	0.14

The topology of the person FSM is also learned in this fashion as any transitions that are never observed place a 0 in the entry for that table. The prior for this model, or hidden state X_1, is set to a uniform probability distribution since CAMEO is uncertain as to what state each person will be in when first viewed.

The observation model $P(Y_t|X_t)$ consists of real-valued vectors of face displacements. Relative displacements are used rather than absolute positions because CAMEO will not be aware of the positions of people in the environment nor will it be aware of its own relative distance and bearing to the people. As such, it is possible for a person in the sit state to have a pixel Y value that is higher than a person

in the stand state. In order to train the models, a set of exemplar observation sequences for the different states were obtained. These were used to generate the appropriate Gaussian (μ, Σ) conditional probability distribution for the observation states, as shown in figure 7. These values were used in the experimental section of the paper.

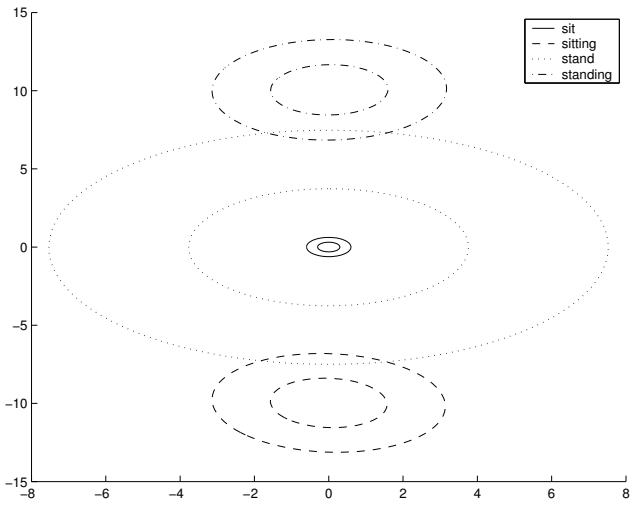


Figure 7. Learned Gaussian distributions for the real-valued observation vectors corresponding to the four possible values of the hidden state. The first and second standard deviations are shown for each distribution.

4.3. Meeting State

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. Let such a meeting model be defined as $M = \{S, T, D\}$, where S is the vector of allowable states, T is the transition matrix between states, and D is a minimum duration for being in that state. The state duration is useful to avoid noise in the model caused by the occasional misclassification of individual person states.

Figure 8 illustrates an example of a simple finite-state machine representation of the a meeting Markov model. In this example, the individual states are classified by the following sets of person states:

Meeting State	Person States
Meeting Start	Everyone in stand state
Presentation	Single person in stand state
General Discussion	Everyone in sit state
Meeting End	Everyone in stand state

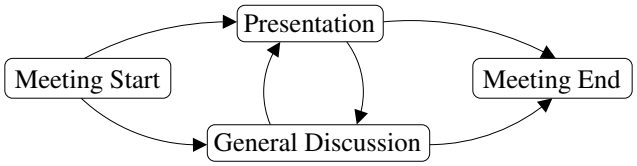


Figure 8. Example finite state machine showing the transitions for the global meeting state.

Inference on this FSM is done simply by computing the set of all the people states and determining whether a state transition should be followed or whether the current state should remain the same. If a state transition should be taken but the minimum time for that state has not been reached, a new state will not be taken.

5. Experimental Results

In order to properly evaluate the DBN classifier in a controlled fashion, we make use of our meeting simulator to generate long test sequences of data. Data-driven simulation as a method for testing multi-agent learning and multi-agent state inference has been used very successfully in the domain of robot soccer [2] where the groups of small-size robots are simulated with very high-fidelity. The use of a simulator allows for careful control of the data so that different instances of events can be produced in any sequence. A simple kinematics engine has been developed which simulates the observed physical positions of people's faces.

5.1. Inferring Person State

Figure 9 shows the raw horizontal and vertical displacements of a person's facial motions generated by the simulator. In this sequence, the person transitions from the sit state to the stand state five different times.

The results of the DBN classification algorithm are shown in Figure 10. Person states are labeled on the vertical axis as follows: sit=1, sitting=2, stand=3, and standing=4. As can be seen from the figure, the classifier was very successful in tracking the person's motion over time. Only 17 states out of all 599 were incorrectly matched. These misclassified states tended to occur

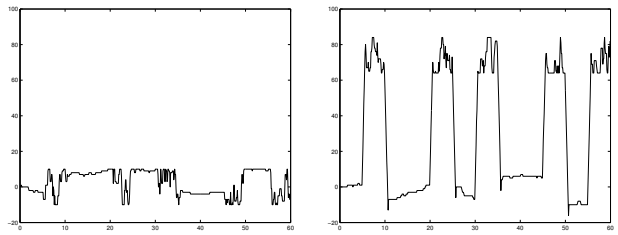


Figure 9. Raw data from one person in the simulated meeting. The top graph shows the horizontal displacement of the face in pixels over time while the bottom shows the vertical displacement over time. The horizontal axis is time in seconds.

on transitions from sit to standing, stand to sitting, and sitting to sit. The classification system was not confused for longer than 0.3 seconds worth of input before it could properly identify the correct state.

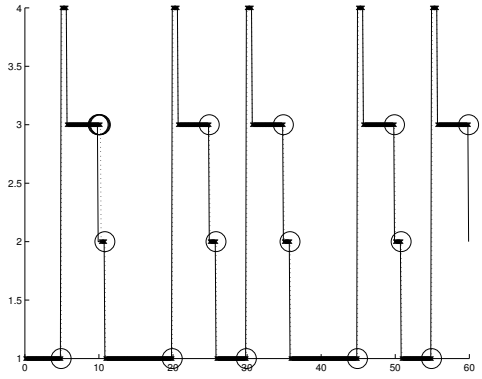


Figure 10. Classified states from a simulated person. State sit=1, sitting=2, stand=3, and standing=4. Circled areas show misclassified states. The horizontal axis is time in seconds. Only 17 states out of 599 were misclassified.

5.2. Inferring Meeting State

The simulator was used to generate data from a meeting that consisted of three people, where one of the people gave a presentation. The presentation was preceded and followed by general discussion. The example Markov model in Figure 8 was used to classify the states of the meeting. Figure 11 shows the results of the meeting state classification. For this simple meeting model, all states were classified correctly.

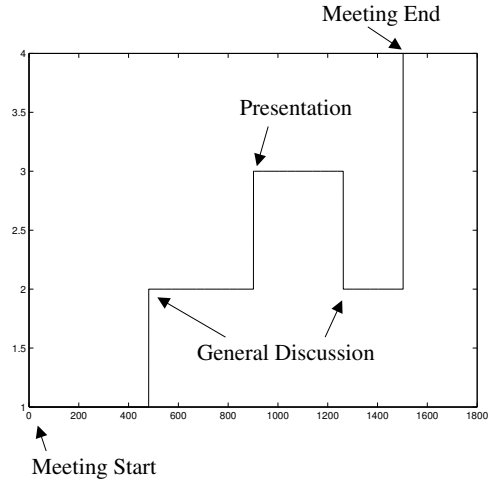


Figure 11. Classified states from a simulated meeting. State Meeting Start=1, General Discussion=2, Presentation=3, and Meeting End=4. The horizontal axis is time in seconds.

6. Summary/Future Work

We are interested in addressing the challenge of developing agent-based technologies to assist users in everyday office-related tasks. These agents need to keep track of their user's state/situation so that they can properly handle day-to-day tasks, and proactively address unexpected events/tasks. We have developed a physical awareness system called CAMEO (Camera Assisted Meeting Event Observer) that is to be used by electronic agents to track the positions of people when they are carrying out group interactions in group meetings. In such situations, individual computer-based interfaces may not be the most appropriate method for an agent to keeping track of its user's interactions. Instead, what is needed is a passive visual observation system such as CAMEO which tracks the positions of people in the environment and infers the state of the individuals as well as the state of the group as a whole.

CAMEO is designed to require minimal room instrumentation and very little calibration to operate. Because very little *a priori* environmental information is expected, CAMEO makes use of a robust facial identification scheme to find and track people's faces in the environment. The motions of the faces are tracked as features and fed into a Dynamic Bayesian Network-based classification system that is

used to infer the person's state. The classifier makes use of a model of human behavior which consists of a finite state machine that is encoded into the Bayesian Network's hidden state's conditional probability distribution. The parameters for the observed states are learned from labeled data. We have shown experimental results from a simulated state model showing how CAMEO is able to infer the state of individual people in the meeting. Finally, high-level contextual information about meeting states is used to define finite state machines that model how a typical meeting progresses. The high-level meeting states are classified by the interactions observed from the individual people models.

For future work, we are actively developing new visual classification algorithms that will augment the existing face detection algorithm. One such improvement is through the use of color histograms for body pose tracking. Once people's faces are identified, a color histogram of the torso is learned that will give CAMEO more information about the person's body motion. This in turn will provide more features to the DBN classifier that will allow it to become more expressive in the kinds of states that can be expressed.

Additionally, the allowable meeting state models need to be expanded to take into account other kinds of meeting formats. A database of different meeting models can be built in order to classify the different kinds of meeting types. In this case, the meeting models would have explicit accept and reject states which would be followed if the person state information didn't match the model correctly.

Acknowledgements

We would like to thank Fernando de la Torre, Raju Patil, Carlos Vallespi, Brett Browning, Takeo Kanade, Betsy Ricker, Johnathan White, and Daniel Hershey 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] J. Allen, H. Kautz, R. Pelavin, and J. Tennenberg. A formal theory of plan recognition and its implementation. In *Reasoning About Plans*, chapter 2, pages 69–126. Morgan Kaufmann Publishers, 1991.

[2] B. Browning and E. Tryzelaar. ÜberSim: A multi-robot simulator for robot soccer. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 948–949, Melbourne, Australia, July 2003.

[3] J. Bruce, M. Bowling, B. Browning, and M. Veloso. Multi-robot team response to a multi-robot opponent team. In *Proceedings of ICRA'03, the 2003 IEEE International Conference on Robotics and Automation*, Taiwan, May 2003, to appear.

[4] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, June 2000.

[5] R. Hamid, Y. Huang, and I. Essa. ARGMode – activity recognition using graphical models. In *Conference on Computer Vision and Pattern Recognition Workshop*, volume 4, pages 38–44, Madison, WI, June 2003.

[6] K. Han and M. Veloso. Automated robot behavior recognition applied to robotic soccer. In J. Hollerbach and D. Koditschek, editors, *Robotics Research: the Ninth International Symposium*, pages 199–204. Springer-Verlag, London, 2000. Also in the Proceedings of IJCAI-99 Workshop on Team Behaviors and Plan Recognition.

[7] P. Hong, M. Turk, and T. S. Huang. Gesture modeling and recognition using finite state machines. In *Proceedings of the Fourth IEEE International Conference and Gesture Recognition*, Grenoble, France, 2000.

[8] G. Kaminka, M. Fidanboylu, A. Chang, and M. Veloso. Learning the sequential coordinated behavior of teams from observations. In *Proceedings of the RoboCup-2002 Symposium*, Fukuoka, Japan, June 2002.

[9] K. Murphy. *Dynamic Bayesian Networks: representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, July 2002.

[10] 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.

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

[12] P. E. Rybski, F. de la Torre, R. Patil, C. Vallespi, M. M. Veloso, and B. Browning. Cameo: Camera assisted meeting event observer. Technical Report TR-04-07, Robotics Institute, Carnegie Mellon University, January 2004.

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

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