A Synthetic Agent System for Bayesian Modeling Human Interactions

Barbara Rosario, Nuria Oliver and Alex Pentland

Vision and Modeling. Media Laboratory MIT,

Cambridge, MA 02139 USA

{rosario,nuria,sandy}@media.mit.edu

Abstract

When building statistical machine learning models from real data one of the most frequently encountered difficulties is the limited amount of training data compared to what is needed by the specific learning architecture. In order to deal with this problem we have developed a synthetic (simulated) agent training system that let us develop flexible prior models for recognizing human interactions in a pedestrian visual surveillance task. We demonstrate the ability to use these prior models to accurately classify real human behaviors and interactions with no additional tuning or training.

1 Introduction

Agent-based solutions have been developed for many different application domains, and field-tested agent systems are steadily increasing in number. Agents are currently being applied in domains as diverse as computer games and interactive cinema, information retrieval and filtering, user interface design, electronic commerce, and industrial process control. In this paper we propose a novel use of agents as prior models and a source of training data for systems that model humans. More specifically our synthetic agents mimic several human interactions that we want to model and recognize.

Our goal is to have a system that will accurately interpret behaviors and interactions within almost any pedestrian scene with little or no train-

1

ing. Our approach to modeling person-to-person interactions is to use supervised statistical learning techniques to teach the system to recognize normal single-person behaviors and common person-to-person interactions. Graphical models [1], such as Hidden Markov Models (HMMs) [2] and Coupled Hidden Markov Models (CHMMs) [3, 4, 5], seem most appropriate for modeling and classifying human behaviors because they offer dynamic time warping, a well-understood training algorithm, and a clear Bayesian semantics for both individual (HMMs) and interacting or coupled (CHMMs) generative processes.

A major problem with a data-driven statistical approach, especially when modeling rare or anomalous behaviors, is the limited number of examples of those behaviors for training the models compared to what is needed by the specific learning architecture. Another critical problem is the generation of models that capture our prior knowledge about human behavior. In order to deal with these problems we have created a synthetic (simulated) agent training system that let us develop flexible prior models for recognizing human interactions in a pedestrian visual surveillance task. Even though the selection of priors is one of the most controversial and open issues in Bayesian inference, we demonstrate the ability to use these prior models to accurately classify real human interactions, with no additional tuning or training. A major emphasis of our work, therefore, is on efficient Bayesian integration of both prior knowledge (by the use of synthetic prior models) with evidence from data (by situation-specific parameter tuning). This approach provides a rather straightforward and flexible technique to the design of priors, one that does not require strong analytical assumptions to be made about the form of the priors[1]. In our experiments we have found that by combining such synthetic priors with limited real data we can easily achieve very high accuracies of recognition of different human-to-human interactions. Thus, our system is robust to cases in which there are only a few examples of a certain behavior or even no examples except synthetically-generated ones.

2   Visual Surveillance System Overview

Our visual surveillance system employs a static camera with wide field-of-view watching a dynamic outdoor scene (the extension to an active camera [6] is straightforward and planned for the next version). A real-time computer

---

[1] Note that our priors have the same form as our posteriors, namely they are Markov models.

vision system segments moving objects from the learned scene. The scene description method allows variations in lighting, weather, etc., to be learned and accurately discounted.

For each moving object (pedestrian) an appearance-based description is generated, allowing it to be tracked though temporary occlusions and multi-object meetings. A Kalman filter tracks the objects location, coarse shape, color pattern, and velocity. This temporally ordered stream of data is then used to obtain a behavioral description of each object, and to detect interactions between objects.

Figure 1 depicts the processing loop and main functional units of our ultimate system (for an extended description of the computer vision system we direct the reader to [7]). (1) The real-time computer vision input module detects and tracks moving objects in the scene, and for each moving object outputs a feature vector describing its motion and heading, and its spatial relationship to all nearby moving objects. (2) These feature vectors constitute the input to stochastic state-based behavior models. Both HMMs and CHMMs, with varying structures depending on the complexity of the behavior, are then used for classifying the perceived behaviors.
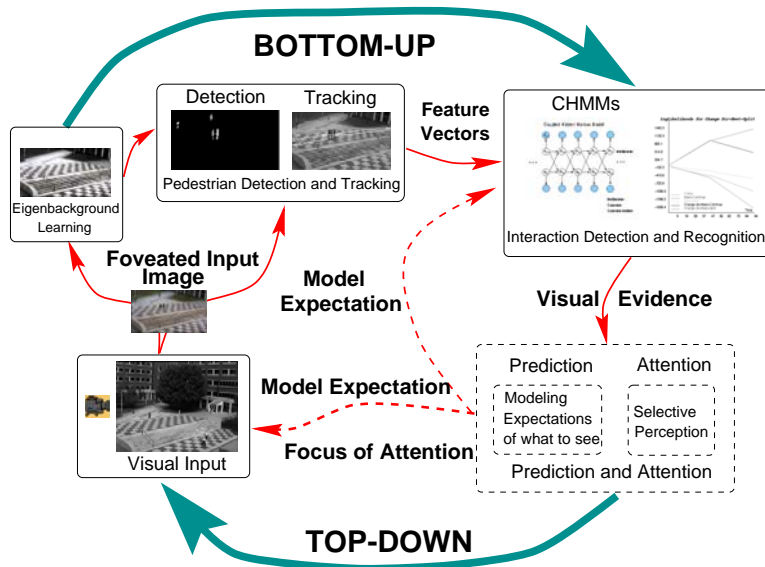

Figure 1: Top-down and bottom-up processing loop

3   Synthetic Behavioral Agents

We have developed a framework for creating synthetic agents that mimic human behavior in a virtual environment. The agents can be assigned different behaviors and they can interact with each other as well. Currently they can generate 5 different interacting behaviors and various kinds of individual behaviors (with no interaction). The parameters of this virtual environment are modeled on the basis of a real pedestrian scene from which we obtained (by hand) measurements of typical pedestrian movement.

3.1   Agent Architecture

Our dynamic multi-agent system consists of some number of agents that perform some specific behavior from a set of possible behaviors. The system starts at time 0, moving discretely forward to time T or until the agents disappear from the scene.

The agents can follow three different paths with two possible directions. They walk with random speeds within an interval; they appear at random instances of time. They can slow down, speed up, stop or change direction independently from the other agents on the scene. When certain preconditions are satisfied a specific interaction between two agents takes place. Each agent has perfect knowledge of the world, including the position of the other agents.

In the following we will describe, without loss of generality, the two-agent system that we used for generating prior models and synthetic data of agents interactions. Each agent makes its own decisions depending on the type of interaction, its location and the location of the other agent on the scene. There is no scripted behavior or a priori knowledge of what kind of interaction, if any, is going to take place. The agents' behavior is determined by the perceived contextual information: relative position of the other agent, speeds, paths they are in, directions of walk, etc., as well as by its own repertoire of possible behaviors and triggering events. For example, if one agent decides to 'follow' the other agent, it will proceed on its own path increasing its speed progressively until reaching the other agent, that will also be walking on the same path. Once the agent has been reached, they will adapt their mutual speeds in order to keep together and continue advancing together until exiting the scene.

4

3.2 Attentional Window: Focus of attention

Given that we are interested in simulating human interactions and that our visual surveillance system is continuously observing the monitored site, there is an attentional mechanism to only record those situations that might eventually lead to an interaction. Specific events signal the system that something interesting might be happening. Whenever a (1) change of direction in one of the agents takes place or (2) the two agents are nearby within the radius of the attentional window, the simulation starts dumping the agents feature vectors. Figure 2 illustrates this mechanism.

Figure 2: Attentional Window in the Synthetic Agents Environment

For each agent the position, orientation and velocity is measured, and from this data a feature vector is constructed which consists of: $\dot{d}_{12}$, the derivative of the relative distance between two agents; $\alpha_{1,2} = sign(< v_1, v_2 >)$, or degree of alignment of the agents, and $v_i = \sqrt{\dot{x}^2 + \dot{y}^2}, i = 1, 2$, the magnitude of their velocities. Note that such feature vector is invariant to the absolute position and direction of the agents and the particular environment they are in.

3.3 Agent Behaviors

The agent behavioral system is structured in a hierarchical way. There are:

1. *Primitive or simple behaviors* that each of the agents performs independently: walk forward, change direction, change speed, stop, start walking.

5

2. At a higher level there are *complex interactive behaviors* to simulate the human interactions. They are composed of a specific temporal succession of simple behaviors. Certain events signal the change of the current simple behavior into another.

The *agents' perceptual system* detects the following events and reacts consequently:

1. Agent close by: when one of the agents falls into the attentional radius of the other agent.

2. Agent reached: when one the agents reaches the other agent such that they occupy roughly the same position.

3. Agent starts moving: after the agents have been talking for a certain amount of time, the 'chatting timer' of at least one of them expires and therefore that particular agent starts moving. This signals the other agent that the conversation is over.

4. Timer expires: Each agent has a timer associated to the stopping behavior, such that they don't stay stopped forever. As soon as one of the agents stops, the timer starts decreasing until it expires. At this moment the agent starts moving again.

In the experiments reported here, we considered five different interacting behaviors that appear illustrated in figures 3,4:

1. Follow, reach and walk together (inter1): The two agents happen to be on the same path walking in the same direction. The agent behind decides that it wants to reach the other. Therefore it speeds up in order to reach the other agent. When this happens it slows down such that they keep walking together with the same speed.

2. Approach, meet and go on separately (inter2): The agents are on the same path but in opposite direction. When they are close enough, if they realize that they 'know' each other, they slow down and finally stop to chat. After talking they go on separately, becoming independent again.

3. Approach, meet and go on together (inter3): In this case, the agents behave like in 'inter2', but now after talking they decide to continue together. One agent changes therefore its direction to follow the other.

6

4. Change direction in order to meet, approach, meet and continue together (inter4): The agents start on different paths. When they are close enough they can see each other and decide to interact. One agent waits for the other to reach it. The other changes direction in order to go toward the waiting agent. Then they meet, chat for some time and decide to go on together.

5. Change direction in order to meet, approach, meet and go on separately (inter5): This interaction is the same as 'inter4' except that when they decide to go on after talking, they separate becoming independent.

Note that in the figures 3, 4 only the portion of the agents' trajectories and feature vectors that fell within the attentional window during the simulation is depicted. We assume that interactions only occur within a certain radius and therefore whenever the agents are too far away there is no record of their trajectories.

Proper design of the interactive behaviors requires the agents to have knowledge about the position of each other as well as synchronization between the successive individual behaviors activated in each of the agents. Figure 5 illustrates the timeline and synchronization of the simple behaviors and events that constitute the interactions.

These interactions can happen at any moment in time and at any location, provided only that the precondititions for the interactions are satisfied. The speeds they walk at, the duration of their chats, the changes of direction, the starting and ending of the actions vary highly. This high variance in the quantitative aspects of the interactions confers robustness to the learned models that tend to capture only the invariant parts of the interactions. The invariance reflects the nature of their interactions and the environment.

4   Behavior Models: HMMs and CHMMs

In this section we describe our framework for building and applying models of individual behaviors and person-to-person interactions. In order to build effective computer models of human behaviors we need to address the question of how knowledge can be mapped onto computation to dynamically deliver consistent interpretations.

Statistical directed acyclic graphs (DAGs) or probabilistic inference networks (PINs) [1, 8] can provide a computationally efficient solution to this

Agent trajectories for inter1   Feature Vector for inter1

Agent trajectories for inter2   Feature Vector for inter2

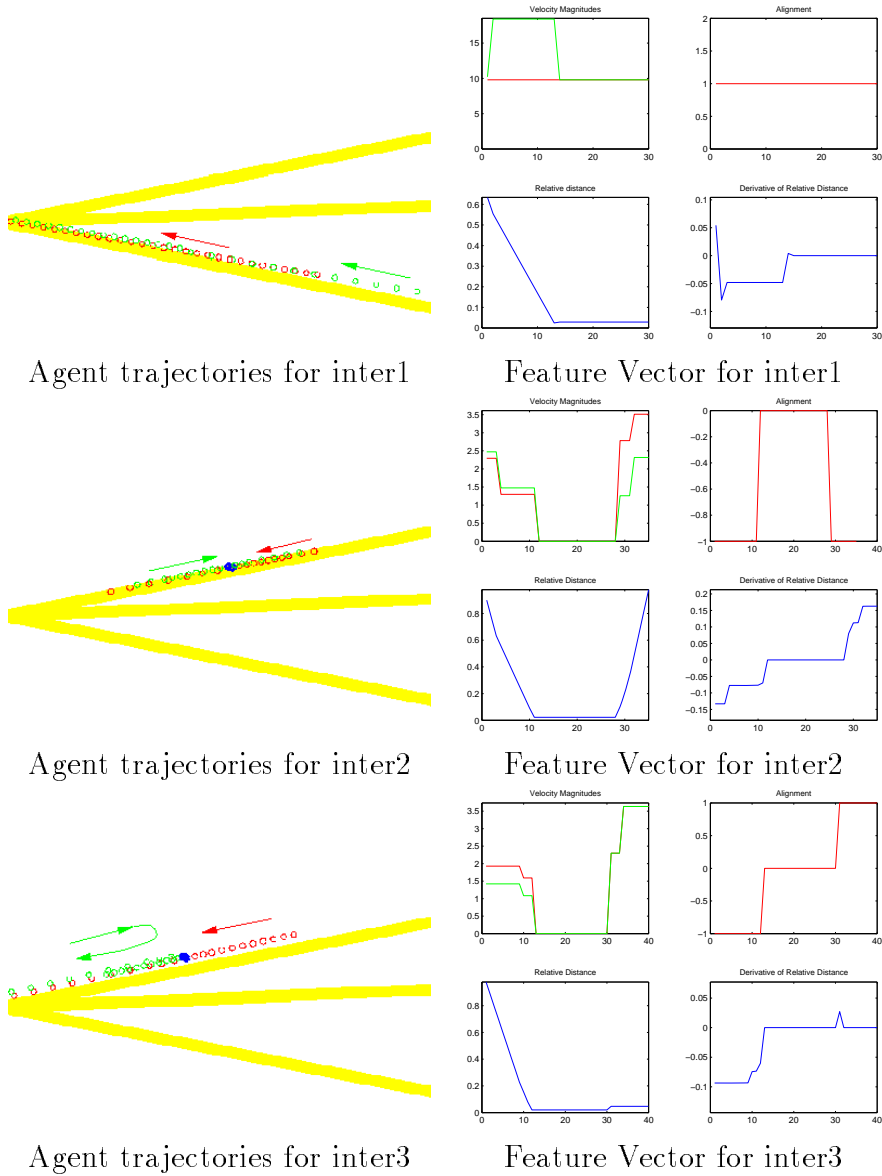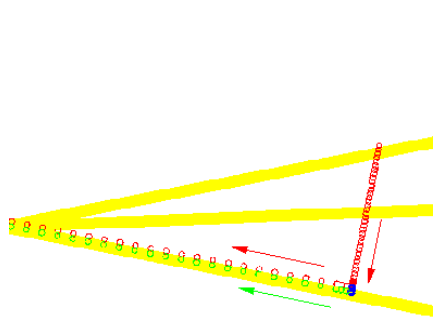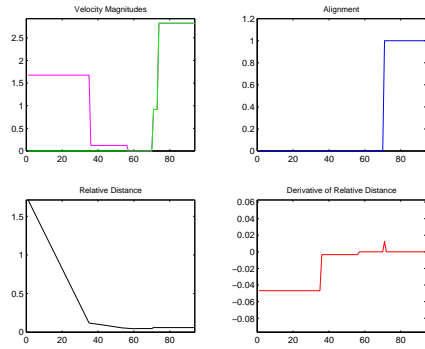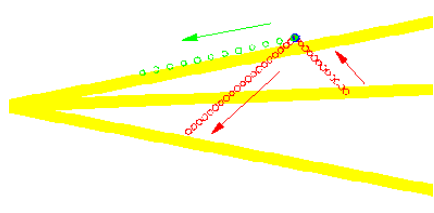Agent trajectories for inter3   Feature Vector for inter3

Figure 3: Example trajectories and feature vector for all the interactions: follow, approach+meet+continue separately, and approach+meet+continue together
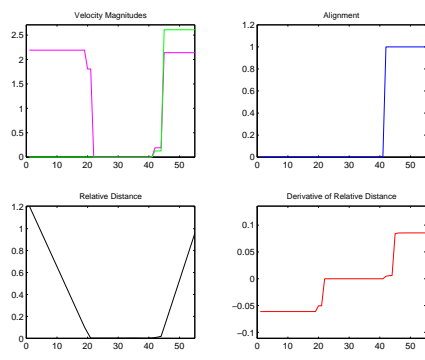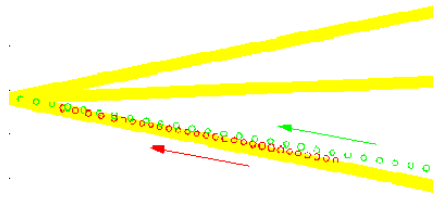
Agent trajectories for inter4

Feature Vector for inter4

Agent trajectories for inter5

Feature Vector for inter5

Agent trajectories for no interaction
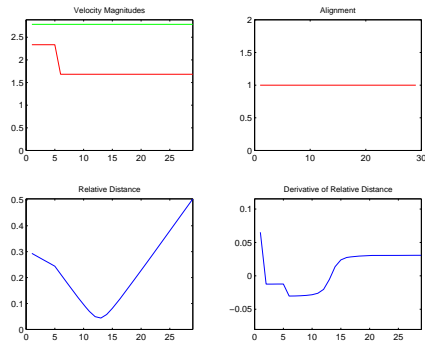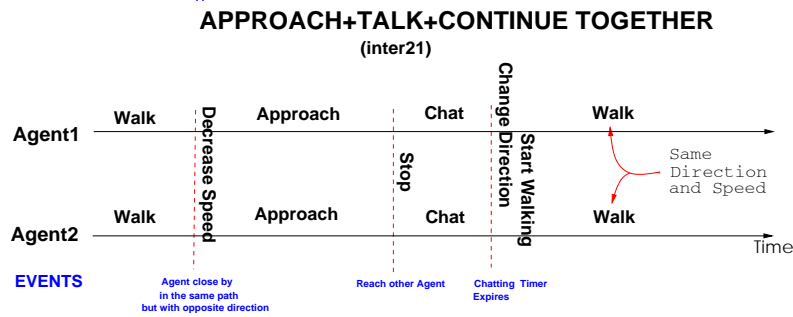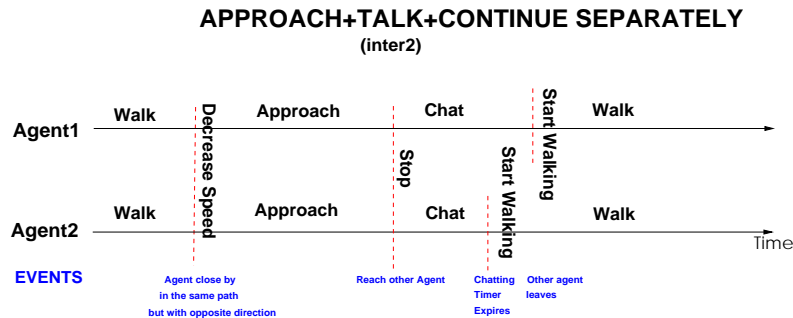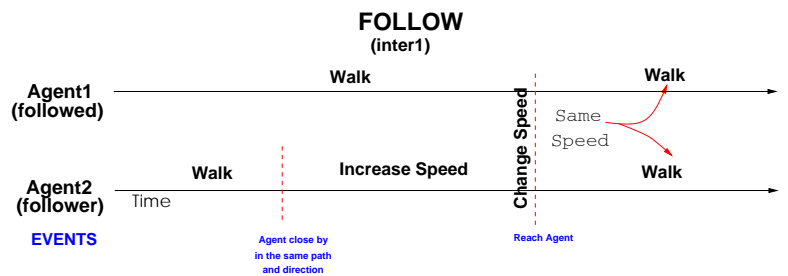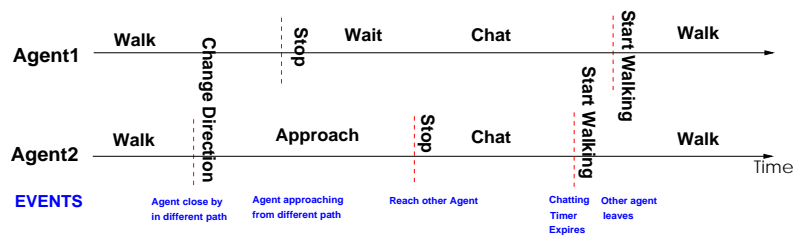
Feature Vector for no interaction

Figure 4: Example trajectories and feature vector for the interactions: change direction+approach+meet+continue separately, change direction+approach+meet+continue together, and no interacting behavior
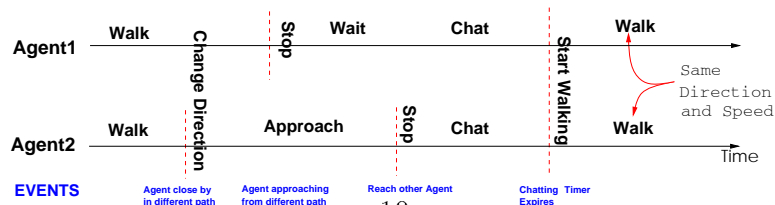
Figure 5: Timeline of the five complex behaviors in terms of events and simple behaviors

Figure 6: A typical image of a pedestrian plaza

question. HMMs and their extensions, such as CHMMs, can be viewed as a particular, simple case of temporal PIN or DAG. In this paper the human behaviors we examine are generated by pedestrians walking in an open outdoor environment. Our goal is to develop a generic, compositional analysis of the observed behaviors in terms of states and transitions between states over time in such a manner that (1) the states correspond to our common sense notions of human behaviors, and (2) they are immediately applicable to a wide range of sites and viewing situations. Figure 6 shows a typical image for our pedestrian scenario.

Hidden Markov models (HMMs) are a popular probabilistic framework for modeling processes that have structure in time. They have a clear Bayesian semantics, efficient algorithms for state and parameter estimation, and they automatically perform dynamic time warping. An HMM is essentially a quantization of a system's configuration space into a small number of discrete states, together with probabilities for transitions between states. A single finite discrete variable indexes the current state of the system. Any information about the history of the process needed for future inferences must be reflected in the current value of this state variable. Graphically HMMs are often depicted 'rolled-out in time' as PINs, such as in figure 7.

However, many interesting systems are composed of multiple interacting processes, and thus merit a compositional representation of two or more variables. This is typically the case for systems that have structure both in time and space. With a single state variable, Markov models are ill-suited to these problems. In order to model these interactions a more complex architecture is needed. We direct the reader to [7] for an extended description
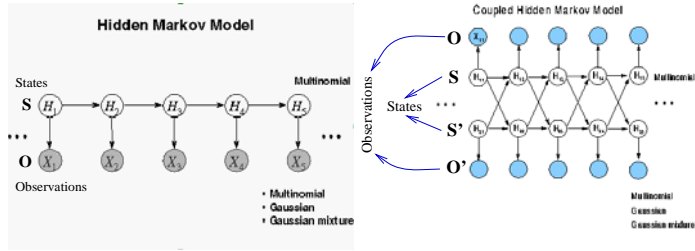
Figure 7: Graphical representation of HMM and CHMM rolled-out in time

of the technical details, performance and comparative analysis of HMMs and CHMMs when modeling human interactions.

Extensions to the basic Markov model generally increase the memory of the system (durational modeling), providing it with compositional state in time. We are interested in systems that have compositional state in *space*, e.g., more than one simultaneous state variable. It is well known that the exact solution of extensions of the basic HMM to 3 or more chains is intractable. In those cases approximation techniques are needed However, it is also known that there exists an exact solution for the case of 2 interacting chains, as it is our case [9, 4].

We therefore use two Coupled Hidden Markov Models (CHMMs) for modeling two interacting processes, in our case they correspond to individual humans. In this architecture state chains are coupled via matrices of conditional probabilities modeling causal (temporal) influences between their hidden state variables. The graphical representation of CHMMs is shown in figure 7. From the graph it can be seen that for each chain, the state at time t depends on the state at time $t - 1$ in both chains. The influence of one chain on the other is through a causal link.

5   Experimental Results

This section describes the experiments we have performed analyzing real pedestrian data using both synthetic and site-specific models (models trained on data from the site being monitored).

12

5.1 Data collection and preprocessing

Using the person detection and tracking system described in [7] we obtained 2D blob features for each person in several hours of video. Up to 20 examples of *following* and various types of *meeting* behaviors were detected and processed.

The feature vector $\bar{x}$ coming from the computer vision processing module is composed of the 2D (x, y) centroid (mean position) of each person's blob, the Kalman Filter state for each instant of time, consisting of $(\hat{x}, \dot{\hat{x}}, \hat{y}, \dot{\hat{y}})$, where $\hat{\ }$ represents the filter estimation, and the (r, g, b) components of the mean of the Gaussian fitted to each blob in color space. The frame-rate of the vision system is of about 20-30 Hz on an SGI R10000 O2 computer. We low-pass filtered the data with a 3Hz cutoff filter and computed for every pair of nearby persons a feature vector consisting of: $\dot{d}_{12}$, derivative of the relative distance between two persons, $|v_i|, i = 1, 2$, norm of the velocity vector for each person, $\alpha = sign(< v_1, v_2 >)$, or degree of alignment of the trajectories of each person. Typical trajectories and feature vectors for an 'approach, meet and continue separately' behavior (interaction 2) are shown in figure 8. This is the same type of behavior as the corresponding 'inter2' displayed in figure 3 for the synthetic agents. Note the similarity of the feature vectors in both cases.
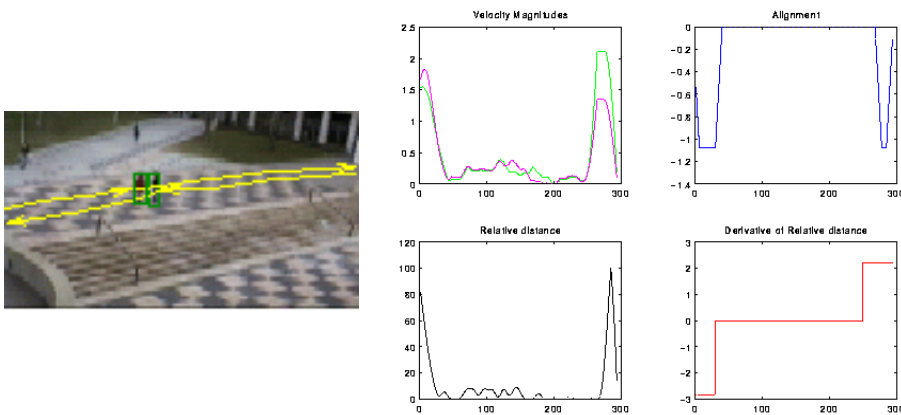

Figure 8: Example trajectories and feature vector for interaction 2, or approach, meet and continue separately behavior.

13

5.2 Human Behavior Models and Evaluation Results

CHMMs were used for modeling three different behaviors: meet and continue together (interaction 3); meet and split (interaction 2) and follow (interaction 1). In addition, an *interaction* versus *no interaction* detection test was also performed. HMMs performed much worse than CHMMs and therefore we omit reporting their results.

In order to evaluate our synthetic agent environment we used CHMMs trained with two types of data:

1. Prior-only (synthetic data) models: that is, the behavior models learned in our synthetic agent environment and then directly applied to the real data with *no additional training or tuning of the parameters.*

2. Posterior (synthetic-plus-real data) models: new behavior models trained by using as starting points the synthetic best models. We used 8 examples of each interaction data from the specific site.
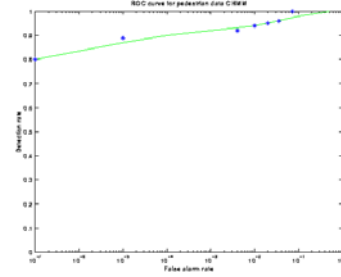
Recognition accuracies for both these 'prior' and 'posterior' CHMMs are summarized in table 9. It is noteworthy that with only 8 training examples, the recognition accuracy on the real data could be raised to 100%. This results demonstrates the ability to accomplish extremely rapid refinement of our behavior models from the initial prior models. Finally the ROC curve for the posterior CHMMs is displayed in figure 9

One of the most interesting results from these experiments is the high accuracy obtained when testing the a priori models obtained from synthetic agent simulations. The fact that a priori models transfer so well to real data demonstrates the robustness of the approach. It shows that with our synthetic agent training system, we can develop models of many different types of behavior — avoiding thus the problem of limited amount of training data — and apply these models to real human behaviors without additional parameter tuning or training.

Agent Parameters sensitivity In order to evaluate the sensitivity of our classification accuracy to variations in the model parameters, we trained a set of models where we changed different parameters of the agents' dynamics by factors of 2.5 and 5. The performance of these altered models turned out to be virtually the same in every case except for the 'inter1' (follow) interaction, which seems to be sensitive to people's relative rates of movement.

14

Testing on real pedestrian data				
	No-inter	Inter1	Inter2	Inter3
Prior CHMMs	90.9	93.7	100	100
Posterior CHMMs	100	100	100	100

Accuracy for real data

ROC curve for real pedestrian data

Figure 9: LEFT TABLE: Accuracy for both untuned, a priori models and site-specific CHMMs tested on real pedestrian data. The first entry in each row is the interaction vs no-interaction accuracy, the remaining entries are classification accuracies between the different interacting behaviors. Interactions are: 'Inter1' follow, reach and walk together; 'Inter2' approach, meet and go on; 'Inter3' approach, meet and continue together. RIGHT FIGURE: ROC curve for real pedestrian data

6   Summary and Conclusions

In this paper we have proposed and evaluated a novel use of agents as prior models and a source of training data for systems that model humans. More specifically our synthetic agents mimic several human interactions that we want to model and recognize. We have described the three main elements of our Bayesian Perceptual System: a synthetic agent environment, a computer vision system and a mathematical modeling framework for recognizing different human behaviors and interactions in a visual surveillance task. Our system combines top-down with bottom-up information in a closed feedback loop, with both components employing a statistical Bayesian approach.

Two different state-based statistical learning architectures, namely HMMs and CHMMs, have been proposed for modeling behaviors and interactions. CHMMs have been found superior to HMMs in terms of both training efficiency and classification accuracy. A synthetic agent training system has been created in order to develop flexible and interpretable prior behavior models, and we have demonstrated the ability to use these a priori models to accurately classify real behaviors with no additional tuning or training. This fact is specially important, given the limited amount of training data available.

Acknowledgments

We would like to sincerely thank Michael Jordan, Tony Jebara and Matthew Brand for their inestimable help and insightful comments.

References

[1] W.L. Buntine, "A guide to the literature on learning probabilistic networks from data.," *IEEE Transactions on Knowledge and Data Engineering*, 1996.

[2] Lawrence R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *PIEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[3] Matthew Brand, Nuria Oliver, and Alex Pentland, "Coupled hidden markov models for complex action recognition," in *In Proceedings of IEEE CVPR97*, 1996.

[4] Matthew Brand, "Coupled hidden markov models for modeling interacting processes," *Submitted to Neural Computation*, November 1996.

[5] N. Oliver, B. Rosario, and A. Pentland, "Graphical models for recognizing human interactions," in *To appear in Proceedings of NIPS98, Denver, Colorado,USA*, November 1998.

[6] R.K. Bajcsy, "Active perception vs. passive perception," in *CVWS85*, 1985, pp. 55–62.

[7] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," in *To appear in Proceedings of ICVS99, Gran Canaria, Spain*, January 1999.

[8] David Heckerman, "A tutorial on learning with bayesian networks," Tech. Rep. MSR-TR-95-06, Microsoft Research, Redmond, Washington, 1995, Revised June 96.

[9] Lawrence K. Saul and Michael I. Jordan, "Boltzmann chains and hidden Markov models," in *NIPS*, Gary Tesauro, David S. Touretzky, and T.K. Leen, Eds., Cambridge, MA, 1995, vol. 7, MITP.