A Vision-Based Architecture for Intent Recognition

Alireza Tavakkoli, Richard Kelley, Christopher King, Mircea Nicolescu, Monica Nicolescu, and George Bebis

Department of Computer Science and Engineering
University of Nevada, Reno, USA
{tavakkol, rkelley, cjking, mircea, monica, bebis}@cse.unr.edu

Abstract. Understanding intent is an important aspect of communication among people and is an essential component of the human cognitive system. This capability is particularly relevant for situations that involve collaboration among multiple agents or detection of situations that can pose a particular threat. We propose an approach that allows a physical robot to detect the intentions of others based on experience acquired through its own sensory-motor abilities. It uses this experience while taking the perspective of the agent whose intent should be recognized. The robot's capability to observe and analyze the current scene employs a novel vision-based technique for target detection and tracking, using a non-parametric recursive modeling approach. Our intent recognition method uses a novel formulation of Hidden Markov Models (HMM's) designed to model a robot's experience and its interaction with the world while performing various actions.

1 Introduction

The ability to understand the intent of others is critical for the success of communication and collaboration between people. The general principle of understanding intentions that we propose in this work is inspired from psychological evidence of a Theory of Mind [1], which states that people have a mechanism for representing, predicting and interpreting each other's actions. This mechanism, based on taking the perspective of others [2], gives people the ability to infer the intentions and goals that underlie action [3]. We base our work on these findings and we take an approach that uses the observer's own learned experience to detect the intentions of the agent or agents it observes. When matched with our own past experiences, these sensory observations become indicative of what our intentions would be in the same situation.

The proposed system models the interactions with the world, acquired from visual information. This information is used in a novel formulation of Hidden Markov Models (HMMs) adapted to suit our needs. The distinguishing feature in our HMMs is that they model not only transitions between discrete states, but also the way in which the parameters encoding the goals of an activity change during its performance. This novel formulation of the HMM representation allows for recognition of the agents' intent well before the actions are finalized.

Our approach is composed of two modules: the Vision module and the HMM module. The vision module performs low-level processing on video frames such as detection and tracking of objects of interest. The detected objects are further processed in the vision module and their 3D positions, distances, and angles are generated. This mid-level information is finally used in the HMM module to perform the two main stages: the activity modeling and the intent recognition.

During the first stage the robot learns corresponding HMM's for each activity it should later recognize. During the intent recognition phase the robot, now an observer, is equipped with the trained HMMs and monitors other agent(s)' performance by evaluating the changes of the same goal parameters, from the perspective of the observed agents.

A significant advantage of the proposed HMM module is that unlike typical approaches to HMMs, which are restricted to be used in the same (training) environment, our models are general and can be transferred to different domains.

The remainder of the paper is structured as follows: Section 2 describes the visual capabilities we developed for this work (vision module). Section 3 summarizes related work in activity modeling and recognition and inferring intent, and presents our novel architecture for understanding intent using HMM's, Section 4 describes our results, and Section 5 summarizes our paper.

2 Vision-Based Perceptual Capabilities

We provide a set of vision-based perceptual capabilities for our system that facilitate the modeling and recognition of actions carried out by the agents. Specifically, we are interested in: detection and tracking of relevant entities, and the estimation of their 3D positions, with respect to the observer.

As the appearance of these agents is generally not known a priori, the only visual cue that can be used for attracting the robot's attention toward them is image motion. Our approach makes significant use of more efficient and reliable techniques traditionally used in real-time surveillance applications, based on background/foreground modeling, structured as follows:

- During the *activity modeling stage*, the robot is moving while performing various activities. The appearance models of the other mobile agents, necessary for tracking, are built in a separate, prior process where the static robot observes each agent that will be used for action learning.
- During the *intent recognition stage*, we assume that the camera is static while the robot observes the actions carried out by the other agents. The static camera allows the use of a foreground-background segmentation technique in order to build the models on-line, and to improve the tracking speed.

2.1 Detection and Tracking

For tracking we use a standard kernel-based approach [4], where the appearance model for each detected region is represented by a histogram-based color dis-

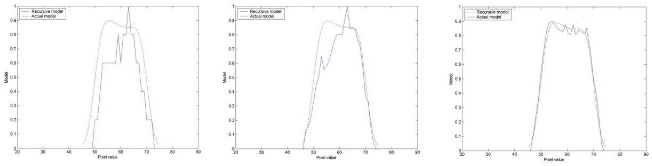


Fig. 1. Model evolution after 10 frames (left) , 50 (middle) and 100 frames (right).

tribution. The detection is achieved by building a representation of the scene background and comparing the new image frames with this representation.

Because of inherent changes in the background, such as fluctuations in monitors and fluorescent lights, waving flags and trees, water surfaces, etc. the background may not be completely stationary. In the presence of these types of backgrounds, referred to as quasi-stationary, more complex background modeling techniques are required.

In parametric background modeling methods, the model is assumed to follow a specific distribution whose parameters must be determined. Mixtures of Gaussians are used in [5]. A Bayesian framework that incorporates spectral, spatio-temporal features to characterize the background is also proposed in [6].

As opposed to this trend, one of the most successful approaches in background modeling [7] proposes a non-parametric model. The background representation is drawn by estimating the probability density function of each pixel by using a kernel density estimation technique.

The Background Model. In this work, we use the *non-parametric modeling*, which estimates the density directly from the data, without any assumptions about the underlying distribution. This avoids having to choose a specific model (that may be incorrect or too restricting) and estimating its distribution parameters. It also addresses the problem of background multi-modality, leading to significant robustness in the case of quasi-stationary backgrounds.

In order to preserve the benefits of non-parametric modeling while addressing its limitations, we propose a *recursive modeling* scheme. Our approach for background modeling employs a recursive formulation, where the background model $\theta_t(x)$ is continuously updated according to equation (1):

$$\hat{\theta}_t(x) = (1-\beta_t) \times \theta_{t-1}(x) + \alpha_t \times H_{\Delta}(x - x_t) \quad : \quad \sum_x \theta_t(x) = 1 \tag{1}$$

The model $\theta_t(x)$ corresponds to a probability density function (distinct for each pixel), defined over the range of possible intensity (or color) values x. After being updated, the model is normalized according to equation (1), so that the function takes values in [0,1], representing the probability for a value x at that pixel to be part of the background. This recursive process takes into consideration the model at the previous image frame, and updates it by using a kernel function (e.g., a Gaussian) $H_{\Delta}(x)$ centered at the new pixel value x_t.

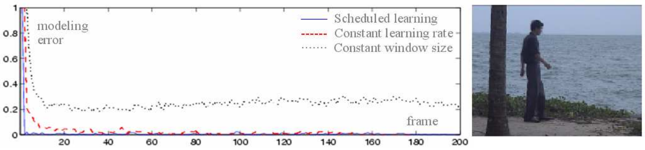


Fig. 2. Convergence speed.

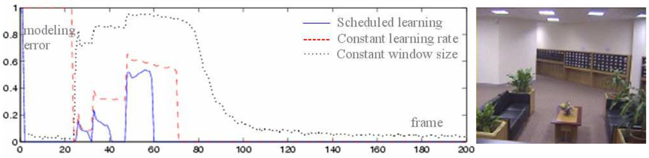


Fig. 3. Recovery speed from sudden global changes.

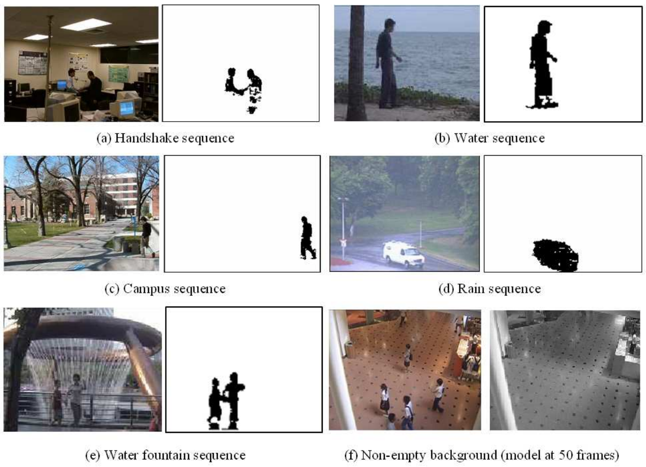
In order to allow for an effective adaptation to changes in the background, we use a *scheduled learning* approach by introducing the learning rate α_t and forgetting rate β_t as weights for the two components in equation (1). The learning and forgetting rates are adjusted online, depending on the variance observed in the past model values. This schedule makes the adaptive learning process converge faster without compromising the stability and memory requirements of the system while successfully handling both gradual and sudden changes in the background independently at each pixel.

Discussion and Results. Fig. 1 shows the updating process using our proposed recursive modeling technique. It can be seen that the trained model (solid line) converges to the actual one (dashed line) as new samples are introduced. The actual model is the probability density function of a randomly generated sample population and the trained model is generated by using the recursive formula presented in equation (1).

Fig. 2 illustrates the convergence speed of our approach with scheduled learning, compared to constant learning and kernel density estimation with constant window size. Fig. 3 compares the same approaches in terms of recovery speed after sudden illumination changes (three different lights switched off in sequence).

Results on several challenging sequences are illustrated in Fig. 4, showing that the proposed methodology is robust to noise, gradual illumination changes, or natural scene variations, such as local fluctuating intensity values dues to monitor flicker (a), waves (b), moving tree branches (c), rain (d), or water motion (e). The ability to correctly model the background even when there are moving objects in every frame is shown in Fig. 4(f).

Quantitative estimation. The performance of our method is evaluated quantitatively on randomly selected samples from different video sequences, taken from [6]. The value used, is the similarity measure between two regions A and B, defined as $S = \frac{A \cap B}{A \cup B}$, where region A corresponds to the detected foreground and B is the actual foreground mask. This measure is monotonically increasing with the similarity of the two masks, with values between 0 and 1.


(a) Handshake sequence

(b) Water sequence

(c) Campus sequence

(d) Rain sequence

(e) Water fountain sequence

(f) Non-empty background (model at 50 frames)

Fig. 4. Background modeling and foreground detection in the presence of quasi-stationary backgrounds.

Table 1. Quantitative evaluation and comparison. The sequences are Meeting Room, Lobby, Campus, Side Walk, Water Surface and Fountain, from left to right.

Videos	MR	LB	CAM	SW	WS	FT	**Avg**
Proposed	0.92	0.87	0.75	0.72	0.89	0.87	**0.84**
Statistical Modeling [6]	0.91	0.71	0.69	0.57	0.85	0.67	**0.74**
Mixture of Gaussians [5]	0.44	0.42	0.48	0.36	0.54	0.66	**0.49**

Table 1 shows the similarity measure for several video sequences where ground truth was available, as analyzed by our method, the mixture of Gaussians [5], and the statistical modeling [6]. It can be seen that the proposed approach clearly outperforms the others, while also producing more consistent results over a wide range of environments. We also emphasize that in the proposed method the thresholds are estimated automatically (and independently at each pixel), and there is no prior assumption needed on the background model.

The scheduled learning scheme achieves a high convergence speed, and a fast recovery from expired models, allowing for successful modeling even for non-empty backgrounds (when there are moving objects in every frame). Its adaptive localized classification leads to automatic training for different scene types and for different locations within the same scene.

2.2 Estimation of 3D Positions

We employ the robot-mounted laser rangefinder for estimating the 3D positions of detected agents with respect to the observing robot. For each such agent, its position is obtained by examining the distance profile from the rangefinder in the direction where the foreground object has been detected by the camera. In order to determine the direction (in camera coordinates) through a pixel, the intrinsic camera parameters are first obtained with an off-line calibration process.

For the intent recognition stage (once the 3D position of each agent is known with respect to the camera) a simple change of coordinates allows the observing robot to take the perspective of any participating agent. This is done in order to map its current observations to those acquired during the action learning stage.

3 General Architecture for Intent Understanding

HMM's are powerful tools for modeling processes that involve temporal sequences and have been successfully used in applications involving speech and sound. Recently, HMMs have been used for activity understanding. They display a significant potential for their use in activity modeling and inferring intent. While some of the existing approaches allude to the potential of using HMMs to learn the user's intentions, these systems fall short of this goal: the approach allows detecting that some goal has been achieved only *after* observing its occurrence. However, for tight collaborative scenarios or for detection of potentially threatening situations, it is of particular importance to detect the intentions *before* the goals of such actions have actually been achieved.

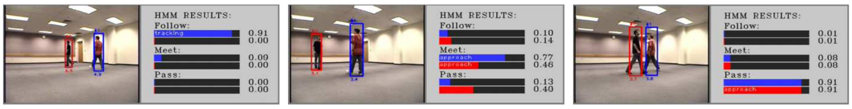
An application of HMMs that is closer to our work is that of detecting abnormal activity. The methods used to achieve this goal typically rely on detecting inconsistencies between the observed activity and a set of pre-existing activity models [8]. Intent recognition has also been addressed from the perspective of intent inference and plan recognition for collaborative dialog [9], but these methods use explicit information in order to infer intentional goals. Our robotic domain relies entirely on implicit cues that come from a robot's sensory capabilities, and thus requires different mechanisms for detecting intent.

3.1 Novel HMM Formulation

Hidden Markov Models have found greatest use in problems that have inherent temporality, to represent processes that have a time-extended evolution. The main contribution of our approach consists in choosing a different method for constructing the model. This HMM formulation models an agent's interaction with the world while performing an activity through the way in which parameters that encode the task goals are changing.

With this representation, the *visible* states reliably encode the changes in task goal parameters while the *hidden states* represent the hidden underlying intent of the performed actions. The reason for choosing the activity goals as the parameters monitored by the HMM is that goals carry intentional meanings.

Activity Modeling. During this stage, the robot uses its experience of performing various activities to train corresponding HMM's. The robot is equipped with a basis set of behaviors and controllers that allow it to execute these tasks. We use a schema-based representation of behaviors, similar to that described in [10]. We experimented with *Following*, *Meeting*, *Passing By*, *Picking Up* and



(a) Follow (b) Meet (c) Pass by

Fig. 5. Intent recognition for different activities.

Dropping off an object. While executing these activities, the robot monitors the changes in the corresponding behaviors' goals.

For a meeting activity, the *angle* and *distance* to the other person are parameters relevant to the goal. The robot's observable symbol alphabet models all possible combinations of changes that can occur: increasing (++), decreasing (−−), constant (==), or unknown (?). The underlying intent of actions is encoded in the HMMs' hidden states.

Repeated execution of a given activity provides the data used to estimate the model transition probabilities a_{ij} and b_{jk} using the Baum-Welch algorithm [11]. During the training stage, the observed, visible states are computed by the observer from its own perspective.

Intent Recognition. The recognition problem consists of inferring, for each observed agent, the intent of the actions they most likely perform from the previously trained HMM's. The observer robot monitors the behavior of all the agents of interest with respect to other agents or locations. Since the observer is now external to the scene, the features need to be computed from the observed agents' perspective rather than from the observer's own point of view. These observations consist of monitoring the same goal parameters that have been used in training the HMM.

For each agent and for all HMM's, the robot computes the likelihood that the sequence of observations has been produced by each model, using the Forward Algorithm [12]. To detect the most probable state that represents the intent of an agent we consider the intentional state emitted only by the model with highest probability. For that model, we then use the Viterbi Algorithm [13] to detect the most probable sequence of hidden states.

4 Experimental Results

To validate our approach we performed experiments with a Pioneer 2DX mobile robot, with an onboard computer, a laser rangefinder and a PTZ Sony camera. The experiments consisted of two stages: the activity modeling phase and the intent recognition phase.

During activity modeling, the robot was initially equipped with controllers for *following*, *meeting* or *passing by* a person for several runs of each of the three activities. The observations gathered from these trials were used to train the HMM's. The goal parameters monitored in order to compute the observable symbols are the distance and angle to the human, from the robot's perspective.

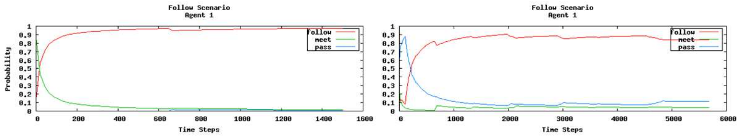


Fig. 6. Model probabilities during two *follow* scenarios.

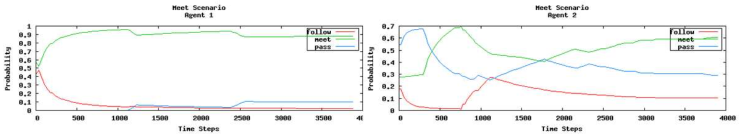


Fig. 7. Model probabilities of two people during the *meet* scenario.

During intent recognition, the robot acted as an observer of activities performed by two people in five different scenarios, which included *following*, *meeting*, *passing by*, and two additional scenarios in which the users switched repeatedly between these three activities. We exposed the robot to different viewpoints of the activities and to show the robustness of the system to varying environmental conditions. The goal of the two complex scenarios is to demonstrate the ability of the system to infer a change in intent as soon as it occurred.

Fig. 5 shows snapshots of the detection and intent recognition for the two runs of each scenario from different viewpoints. The blue and red bars correspond to the blue and red-tracked agent, respectively, whose length represent the cumulative likelihood of the models up to that point in time.

Fig. 6 through Fig. 8 show that the robot is able to infer the correct intent for the *following*, *meeting*, and *passing by* scenarios: the probability for the correct model rapidly exceeds the other models, which have very low likelihoods.

For the *following* scenarios (Fig. 6), we only present the intent of the person who is performing the action. For the other scenarios (Fig. 7 and Fig. 8), we show the intent of both people involved in the activities: the robot is able to detect that both have similar intentions, either related to meeting or passing by. During the complex scenarios the system was capable of adapting to changes in people's activities quickly, and of detecting the correct intentional state of the agents, as shown in Fig. 10.

After these experiments were performed we added two new activity models to the robot's set of capabilities, for *picking up* and *dropping off* objects. Fig. 9 presents the model probabilities of *drop off* and *pick up* activities, respectively.

To provide a quantitative evaluation of our method we analyze the *accuracy rate*, *early detection* and *correct duration*, typically used in HMM's [14]:

Accuracy rate = the ratio of the number of observation sequences, of which the winning intentional state or activity matches the ground truth, to the total number of test sequences.

Early detection (ED) $= t/T$, where T is the observation length and $t^* = min\{t|Pr(\text{winning intentional activity) is highest from time } t \text{ to } T\}$.

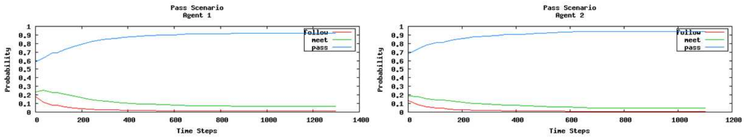
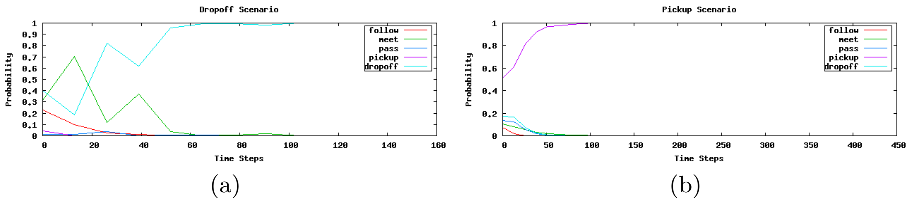


Fig. 8. Model probabilities of two people during the *passing by* scenario.



(a)

(b)

Fig. 9. Model probabilities during: (a) *drop off* and (b) *pick up* scenarios.

Correct duration (CD) $= C/T$, where C is the total time during which the intentional state with the highest probability matches the ground truth.

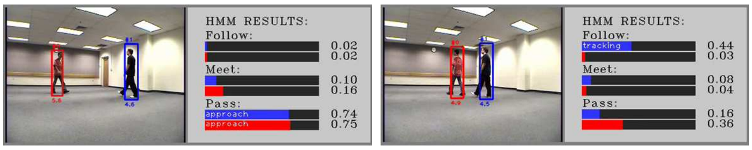
For reliable recognition, the system should have a high *accuracy rate*, a small value for *early detection* and high *correct duration*. The accuracy rate of our system is 100%: all intent recognition scenarios have been correctly identified. Table 2 shows the *early detection* and the *correct duration* for these experiments. The worse results occurred when inferring agent 2's intent, during the meeting scenarios. From our analysis of the data we observed that this result is due to small variations in computing the observable symbols from agent 2's perspective and the high similarity between *meeting* and *passing by*.

5 Conclusion and Future Work

In this paper, we proposed an approach for detecting intent from visual information. We developed a vision-based technique for target detection and tracking that uses a new non-parametric recursive modeling approach. We proposed a novel formulation of Hidden Markov Models (HMMs) to encode a robot's experiences and its interactions with the world when performing various actions. These models are used through taking the perspective to infer the intent of other agents before their actions are finalized. This is in contrast to other activity recognition approaches which only detect an activity , after it is completed. We validated this architecture with an embedded robot, detecting the intent of people performing multiple activities. We are working on expanding the repertoire of activities for the robot to more complex navigation scenarios.

References

1. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? In: Behavioral and Brain Sciences. Volume 1. (1978) 515–526



(a) Red and blue pass by

(b) Blue follows red

Fig. 10. Results from complex scenarios.

Table 2. Quantitative evaluation.

Scenario	Follow	Meet		Pass		Drop off	Pick up
	Both	Agent1	Agent 2	Agent 1	Agent 2	Both	Both
Avg. ED[%]	2.465	4.12	49.85	0	0	11.53	0
Avg. CD[%]	97.535	95.88	66.82	100	100	90.38	100

2. Gopnick, A., Moore, A.: Changing your views: How understanding visual perception can lead to a new theory of mind. In: Children's Early Understanding of Mind. (1994) 157–181
3. Baldwin, D., Baird, J.: Discerning intentions in dynamic human action. In: Trends in Cognitive Sciences. Volume 5. (2001) 171–178
4. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. In: IEEE Transactions on Pattern Analysis and Ma-chine Intelligence. Volume 25. (2003) 564–577
5. C. Stauffer, a.W.G.: Learning patterns of activity using real-time tracking. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 22. (2000) 747–757
6. Li, L., Huang, W., Gu, I., Tian, Q.: Statistical modeling of complex backgrounds for foreground object detection. In: IEEE Transactions on Image Processing,. Volume 23. (2004) 1459–1472
7. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.: Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. In: Proceedings of the IEEE. Volume 90. (2002) 1151–1163
8. Duong, T., H. Bui, a.D.P., Venkatesh, S.: Activity recognition and abnormality detection with the switching hidden semi-markov model. In: IEEE Intl. Conference on Computer Vision and Pattern Recognition. (2005)
9. Grosz, B.J., Sidner, C.L.: Plans for discourse. In: Intentions in communication. (1990) 417–444
10. Arkin, R.C.: Behavior-based robotics. (1998) 417–444
11. Baum, L.E., Peterie, T., Souled, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. In: Ann. Math. Statist. (Volume 41.) 164–171
12. Rabiner, L.R.: A tutorial on hidden-markov models and selected applications in speech recognition. In: Proceedings of the IEEE. (Volume 77.)
13. Jr., G.D.F.: The viterbi algorithm. In: Proceedings of the IEEE. (Volume 61.) 2268–278
14. Nguyen, N., Phung, D., Venkatesh, S., Bui, H.: Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. (In: IEEE Intl. Conference on Computer Vision and Pattern Recognition) 955–960