Adaptive Agent Based System for Knowledge Discovery in Visual Information

Ph.D. Thesis Proposal
Alvaro Soto

The Robotics Institute
Carnegie Mellon University

August-2000

Abstract

One of the main limitations of most current machine vision systems is a lack of flexibility to consider the wide variety of information provided by visual data. The research proposed here aims to improve this situation by the development of an adaptive visual system able to selectively combine information from different visual algorithms.

The problem is cast as a knowledge discovery problem, where the two main steps are detection and characterization of relevant patterns. The algorithms will be able to perceive different attributes of the visual space such as color, depth, motion or specific shapes. The intended system should be able to adaptively select and combine the information provided by the algorithms according to the quality of the information given by each of them.

The system proposed is based on an *intelligent agent*[1] paradigm. Each visual module will be implemented as an agent that will be able to adapt its behavior according to the relevant task and environment constraints. The adaptation will be provided by a local self-evaluation function on each agent. Cooperation among the agents will be given by a probabilistic scheme that will integrate the evidential information provided by them.

The proposed system aims to achieve two highly desirable attributes of an engineering system: *robustness* and *efficiency*. By combining the outputs of multiple vision modules the assumptions and constrains of each module will be factored out to result in a more robust system overall. Efficiency will be still kept through the on-line selection and specialization of the algorithms according to the relevant structures and conditions present at each time in the visual scene.

The advantages of the approach proposed here will be demonstrated in two frequent problems faced by a mobile robot: dynamic target tracking and obstacle detection.

[1] For a definition of an *intelligent agent* see section 2.2

1. INTRODUCTION

• *Knowledge Discovery*

Today technology is changing the way we produce and handle information. The increasing capabilities and falling cost of suitable equipment to acquire, process, and store information is allowing its massive use. In this new technological context large sources of information of diverse types are becoming increasingly available.

While the opportunities of an effective use of this information are enormous, there is still a lack of effective tools that can fully exploit its potential. From the seemingly endless information paths of Internet to the case of a mobile robot collecting information from its environment, there is an increasing need for the development of automatic tools able to transform the information available in useful knowledge. Manual analysis is not longer a viable solution, and automated knowledge discovery has emerged as the key technology that can take advantage of the new massive sources of data.

In general, the process of automatic knowledge discovery is given by the extraction of regularities or patterns in the information data. The identification of regular structures in the data allows making predictions and generalizations, justifying the knowledge acquisition through an inductive learning step. In contrast to information theory where the relevant part of the information lays in the unpredictable part of a signal, in the case of knowledge discovery the relevant information lays in the redundancy of non-accidental structures.

• *Robotics Domain*

In the Robotics domain the problem of knowledge discovery from sensing information is highly relevant. While today it is possible to equip a robot with many sensors and sophisticated locomotion capabilities, the perception skills of most robots are still rather limited.

In order to move robots out of labs to perform useful tasks in natural environments, it is needed to equip them with more powerful perception systems able to acquire useful knowledge from diverse sources of information. Today the main challenge for robots is not the controllability but the observability problem.

• *Visual Perception*

In particular, the case of visual perception is a very attractive option to equip a robot with suitable perceptual capabilities. In contrast to other sensor modalities, vision can allow to perceive a large number of different features of the environment such as color, shape, depth, motion, and so on. Unfortunately, visual perception has turned to be a complex problem, even though there have been advances in the field, still it is not possible to find a reliable vision system able to safely guide the way of a robot on an unstructured and dynamic environment.

The main problem is that a dynamic unconstrained environment presents numerous ambiguities to a visual perception system. Patterns, tendencies, and models lay in a

complex high dimensional space of shapes, distances, colors, sounds, past experiences, and so on. Different visual attributes form a multidimensional space where the number of subspaces that can contain relevant features grows exponentially with the number of dimensions. This stresses the need for good heuristics or previous high-level knowledge to bias the search and to keep the problem manageable.

Most of current successful applications in machine vision have heavily used task and environment constraints to cope with the high dimensionality and inherent ambiguity of visual information. The typical approach relies on simplifications of the environment or on good engineering work to identify relevant visual attributes that allow solving a specific visual task. For example, consider the case of a robot localization system based on artificial landmarks. In this case, previous knowledge about the structure of the landmarks provides strong constraints that allow constructing algorithms especially designed to detect key visual attributes [1][2]. In the same way, recent successful vision systems able to detect people or cars are examples of specific visual applications where good engineering work allows an off-line identification of key visual attributes to complete the task [3][4].

• *Adaptation*

The main problem with the previous approach is lack of flexibility. In general, it is difficult to know a priori which part of the visual space and which set of visual attributes will convey enough information to extract the knowledge needed to complete a task. Problems such as partial occlusion, changes in illumination, or different postures constantly modify the quantity of information or entropy of the different visual attributes. As a consequence, there is a high variability about the adequate set of attributes to complete a given task.

As an example consider ALVINN [5], a perceptual visual system designed to steer a car in natural environments using a neural net learning algorithm. After training, the main internal features learned by ALVINN were the edges of the road. With this knowledge ALVINN was able to demonstrate a reasonable performance, but it irremediable failed in situations where the edges of the road were obstructed by other passing vehicles, or were missing as on bridges or crossing points. The main problem with ALVINN was its lack of adaptability to use alternative sources of information such as centerlines, other traffic, roadway signs, and so on.

In contrast to ALVINN human drivers are remarkable robust to changes in the driving conditions. This great robustness of the human visual system can be explained by its extreme flexibility to adapt to the changing conditions of the environment by selecting appropriate sources of information.

The lack of flexibility of most actual machine vision systems to consider alternatives sources of information has limited their robustness to successfully operate in natural environments, being one of the main reasons to prevent the more extensive use of these types of systems.

• *Information Theory and Probabilistic Reasoning*

The case of ALVINN illustrates the potential benefits of adding adaptability to a visual perception system. It also shows that the level of adaptation should be directly related to the degree of useful knowledge that a perception system can be able to extract from the different sources of information. This introduces the need for appropriate metrics that can evaluate the quantity of information conveyed by different subsets of visual attributes. Information theory provides such metrics through concepts such as entropy or information gain.

Another important intrinsic feature of most information worlds is uncertainty. In order to create robust systems, it is highly desirable to use a representation able to characterize the ambiguity inherent to natural scenarios. Probability theory provides a solid mathematical framework to represent and to reason under uncertainty. In particular, Bayesian inference provides a suitable framework to combine knowledge from different sources of information.

I foresee that the synergistic interaction between elements from probabilistic reasoning and information theory can be a powerful combination to implement a knowledge discovery system. The intuition behind this idea lays at the heart of two highly desirable attributes of an engineering system: *robustness* and *efficiency*. In order to achieve robustness one needs to use a representation such as the one provides by probability theory that allows to reason under ambiguity. Unfortunally, in many situations the combinatorial explosion in the number of possible hypotheses or explanations makes a full probabilistic approach impractical. This is especially true for high dimensional and real time applications such as the one intended in this work. So, one also needs to be efficient, and in order to achieve efficiency one needs to use tools able to quantify ambiguity, tools that leads the inference engine to the more prominent hypothesis and sources of information. Information theory can provide such tools.

• *Proposed work*

The focus of the dissertation research I propose to perform is the development of an adaptive visual system able to selectively combine information from different visual algorithms. These algorithms will be able to perceive different attributes of the visual space such as color, depth, motion or specific shapes.

The intended system will be based on an *intelligent agent*[2] paradigm. Each visual module will be implemented as an agent that will be able to adapt its behavior according to the relevant task and environment constraints. The adaptation will be provided by a local self-evaluation function on each agent. Cooperation among the agents will be given by a probabilistic scheme that will integrate the evidential information provided by them.

Using the power of probability theory for representing and reasoning under uncertainty, and elements from information theory to lead the inference engine to prominent

[2] For a definition of an *intelligent agent* see section 2.2

hypothesis and information sources, the proposed system aims to achieve two highly desirable attributes of an engineering system: *robustness* and *efficiency.*

By combining the outputs of multiple vision modules the assumption and constrains of each module will be factored out to result in a more robust system overall. Efficiency will be still kept through the on-line selection and specialization of the algorithms according to the relevant structures and conditions present at each time in the visual scene.

The adaptive cooperation of diverse visual algorithms will provide great flexibility about the type of visual structures and therefore the kind of knowledge that can be extracted from visual information. Although the system is presented for the case of visual information, the ideas can be extended to other domains that perform unsupervised knowledge extraction from dynamic multidimensional information sources.

The research proposed in this work is particularly relevant for the case of dynamic visual tasks with a high variability about the subsets of visual attributes that can characterize relevant visual structures. This includes visual tasks such as dynamic target tracking, obstacle detection, and identification of landmarks in natural scenes. In particular, the advantages of the approach proposed here will be demonstrated in two frequent problems faced by a mobile robot: dynamic target tracking and obstacle detection. This document includes preliminary results in this direction for both visual problems.

2. PROPOSED RESEARCH

2.1. Detectors and Specialists

In most visual applications one is interested in the detection of some type of visual structures or patterns. Usually these visual structures define regularities in the visual space, which can be characterized by a combination of visual properties, such as spatial continuity, coherent motion, specific shape and so on. This characterization of visual structures using their more prominent visual properties is which allow their posterior detection.

In general given the great variability of most natural scenarios, it is difficult to know a priori which part of the visual space and which set of visual attributes will define a relevant visual structure. This presents a duality between detection (where is it?) and characterization (how is it?). Knowledge about the location of a relevant visual structure in an input image simplifies the problem of finding the more appropriate set of visual attributes to build appearance model for the pattern. In the same way, knowledge about the more appropriate set of visual attributes to characterize a relevant visual structure simplifies the problem of detecting the pattern in an input image.

This idea suggests a two-step approach to find relevant structures in a complex visual space. First one needs general tools able to explore the incoming information looking for possible structures without relying in specific illumination, views, posture, etc. Then once possible candidate structures have been identified, it is possible to use more specific tools

able to look for specific supporting evidence in order to provide more efficient and robust appearance models. I will refer to these steps as detection and specialization, and the respective algorithms as *detectors* and *specialists*.

The intuition behind this idea is that detectors and specialists use different types of constraints. While in natural environments the detection of visual structures allows only the use of very general weak constraints, after a relevant structure has been detected it is possible to select more appropriated and specific constraints that allow a more robust and efficient characterization of a given visual structure. This more effective characterization is especially useful for the case of dynamic visual sequences, where the visual system needs to keep track of the structures in time.

For example if the goal of a given visual system is to track people, one possible approach could use a *detector* based on depth continuity and shape, and a set of *specialists* based on motion, color, position, and other visual attributes. At the beginning of the tracking the lack of specific knowledge about the types of colors that a given person is wearing, as long as its position or its type of motion will not allow to use these types of visual information for the tracking task, so the system would have to relies in the more general detector base on depth continuity and shape. Now, after a person is detected, the set of *specialists* can provide a more adequate appearance model, specially tuned to the more prominent visual attributes of the intended target. Giving that tracking based only on depth continuity and shape will inevitable fail when a detected person passes close to another person or a bulky object, the on-line specialization of the tracker using the strong constraint providing by the *specialists* will allow a more robust tracking.

2.2. Intelligent Agents

Even though there are a diversity of views about what intelligent agents are, there is a general agreement that the main features that distinguish an intelligent agent are *autonomy*, *adaptation* and *sociability* [6]. Autonomy provides the independency that allows the agent to exhibit an opportunistic behavior in agreement with its goals. Adaptation provides the flexibility that allows the agent to change its behavior according to the conditions of the environment. Sociability provides the communication skills that allow the agent to interact with other artificial agents and humans.

This work makes use of multiple agents that can simultaneously analyze different properties of the incoming information. These agents will act as a group of experts where each agent will have a specific knowledge area. This scheme provides a high degree of abstraction and modularity, which facilitate the design and scalability of the system.

The system will consist of two main types of agents. A set of agents known as *detectors* will have the mission to explore the incoming information providing hypothesis about the location of possible relevant structures. The main requirement for a *detector agent* will be the capacity to provide useful information under different conditions. For example, in the case of visual information a *detector agent* should be able to detect structures under different illumination, views or posture conditions. A second type of agents will be the *specialists*. These agents will have the mission to adapt their behavior according to the

type of structures detected. This adaptation will be possible by the specialization of *detector agents* or through the opportunistic pro-activation of specific algorithms that encapsulate particular constraints.

All these agents will be able to run in parallel in local or remote machines using a distributed multi-threated software architecture. This will provide the degree of *autonomy* needed by the agent in order to show an opportunistic behavior. Furthermore, each *detector* and *specialist* will be provided with a local self-evaluation function that will be used by the agent to evaluate its own performance. This self-evaluation will be the key element used by the agents to activate *adaptation* mechanisms. Finally, *sociability* among the agents will be given through the integration of information using Bayesian reasoning, which is the topic of the next section.

2.3. Probabilistic Inference through Bayesian Reasoning

- *Bayes' Rule*

Bayesian theory provides a solid mathematical framework for reasoning under uncertainty. Using the language of probability theory, a Bayesian approach provides mechanisms to combine information in order to reason about different hypothetical solutions to a problem. The basic idea is to use the information available for building a probability distribution, which characterizes the relative likelihood of each hypothesis.

The core of the Bayesian technique is the so-called Bayes' Rule :

$$P(h/e) = \frac{P(e/h)*P(h)}{P(e)} = \alpha * P(e/h)*P(h) \qquad (1)$$

P(h / e) is called the posterior conditional probability and represents the probability of a hypothesis h given the information or evidence available e. P(e / h) is called the likelihood function and represents the degree of fitness between the hypothesis and the data. P(h) is called the prior probability and represents the previous belief about the feasibility of each hypothesis h. Finally, as it is stated in (2), P(e) acts as a normalization factor that can be derived from the other terms.

$$P(e) = \sum_{h} P(e/h)*P(h) = 1/\alpha \qquad (2)$$

One of the more suitable features of the Bayesian approach is the decoupling between evidence and previous knowledge given by the likelihood and the a priori terms in (1). This explicitly shows how the inference engine combines the different types of information, being a great help to model the system. The likelihood term allows measuring the support of the incoming information to each of the possible hypothesis. The a priori term allows considering past experiences and task constraints. This is closely related to the theory of regularization commonly used in computer vision. Regularization theory calls for minimizing a cost function that has a data and a regularizer term. The data term considers the evidence while the regularizer term constraints the set of possible hypothesis allowing to solve ill-posed problems.

One common extension to the traditional Bayes's Rule is to account for time, which is a highly relevant dimension for the case of dynamic visual scenes. There is a straightforward method to extend (1) to the dynamic case. Consider the posterior conditional probability distribution for hypothesis *h* at time instant *t* called h_t, given all the evidence $\vec{e}_t$ accumulated until time *t*. Using Bayes's rule this can be expressed as:

$$P(h_t / \vec{e_t}) = \beta * P(e_t / h_t, \vec{e_{t-1}}) * P(h_t / \vec{e_{t-1}}) \qquad (3)$$

Now, assuming that the current evidence e_t can be totally explained by the current hypothesis h_t, and that the dynamic of the system follows a first order Markov process, it is possible to obtain (6) which is the standard way to perform Bayesian inference for the dynamic case.

$$P(h_t / \vec{e_t}) = \beta * P(e_t / h_t) * P(h_t / \vec{e_{t-1}}) \qquad (4)$$

$$P(h_t / \vec{e_t}) = \beta * P(e_t / h_t) * \sum_{h_{t-1}} P(h_t / h_{t-1}, \vec{e_{t-1}}) * P(h_{t-1} / \vec{e_{t-1}}) \qquad (5)$$

$$P(h_t / \vec{e_t}) = \beta * P(e_t / h_t) * \sum_{h_{t-1}} P(h_t / h_{t-1}) * P(h_{t-1} / \vec{e_{t-1}}) \qquad (6)$$

Equation (6) is the equivalent of Bayes' Rule for the time variant case. The posterior density at time *t-1*, $P(h_{t-1} / \vec{e_{t-1}})$ is propagated in time using the dynamics of the system, $P(h_t / h_{t-1})$, to form the new priors for the new time instant. These priors are then multiplied by a likelihood function $P(e_t / h_t)$ in the usual Bayesian fashion to obtain the new posterior $P(h_t / \vec{e_t})$.

• *Estimation of Probability Density Functions*

One practical difficulty using Bayesian modeling is to find adequate probability density functions (pdfs) to represent the different terms required by the Bayes' Rule. The problem is even more difficult for the dynamic case, where one also needs to propagate the pdfs in time. The theory of probability provides several techniques to estimate probability models from data. These techniques are mainly classified in parametric, semi-parametric and non-parametric estimation methods. Two of these estimation techniques are relevant to the system proposed in this work: parametric estimation using Gaussian densities and non-parametric estimation based on stochastic sampling.

A Gaussian density provides a suitable tool to probabilistically model the performance of *detectors* and *specialists* for cases when in average they provide non-bias solutions close to the correct hypothesis. There are several advantages in using Gaussian densities. First, they provide a closed form representation where the parameters can be estimated using

training data and maximum likelihood. Also, the probabilistic models can be easily extended to higher dimensions using multivariate Gaussian densities. Furthermore in the case that the dynamics of the process is lineal and the system noise is Gaussian, the pdfs can be propagated in time using the well-known Kalman Filter [7].

Unfortunately in some cases *detectors* and *specialists* are no so well behaved, in the sense described above. Ambiguous situations can confuse *detectors* and *specialists* producing less predictable results. In this case the probability distribution can have a complex multi-modal shape that cannot be accurately modeled by a Gaussian density. Stochastic sampling provides an alternative estimation approach for these cases.

In stochastic sampling a pdf is represented through a set of samples, each with an associated weight representing its probability. There are several algorithms available to estimate the samples. Likelihood Weighting [8] and Factored Sampling [9] are two of those algorithms.

- *Likelihood Weighting*

Consider a case where there are available a set of hypotheses, a set of observations, and a metric to evaluate the degree of fitness between hypotheses and observations. An intuitive idea to obtain a probabilistic model over the set of hypotheses is to normalize the fitness between hypotheses and observations, and then use this metric as the probability of each hypothesis. In the case of an extremely large hypotheses space, the complete pdf could be approximated using a set of regularly spaced samples.

The previous approach is the basis of the Likelihood Weighting algorithm, which has been widely used for stochastic simulation. The algorithm can be extended to the time variant case using the dynamic of the process to propagate each hypothesis in time and then repeat the weighting procedure using the new set of hypotheses and observations.

- *Factored Sampling*

Although Likelihood Weighting provides a suitable way to represent and to propagate complex pdfs in time, it does not provide a mechanism to allocate the samples in an efficient way. The initial set of samples is allocated in a uniform fashion without considering critical areas of the probability distribution. Factored Sampling overcomes this limitation for the case of a posterior density factored according to (1).

As opposed to Likelihood Weighting, Factored Sampling uses the current beliefs to obtain a more suitable set of samples. Instead of just obtaining the set of samples using a uniform scheme, the samples are obtained by sampling from the current prior distribution. This is particularly important for the case of dynamic inference because it allows using the information gathered so far to obtain a more adequate allocation of the samples.

Although originally the Factored Sampling algorithm was presented for the static case of equation (1) [9], it can be easily extended to the dynamic case of equation (4) [11]. In this

case the Factored Sampling is also called the Condensation algorithm, and it operates in the following way:

First obtain a set of *n* sample hypotheses h_i from $P(h_t / \vec{e}_{t-1})$ or equivalently $\sum_{h_{t-1}} P(h_t / h_{t-1}) * P(h_{t-1} / \vec{e}_{t-1})$, and then weight each of these sample hypotheses h_i by π_i. As it is stated in (7), π_i is a normalized version of $P(e_t / h_i)$ with e_t the current observation.

$$\pi_i = \frac{P(e_t / h_i)}{\sum_{h_i} P(e_t / h_i)} \qquad (7)$$

The set of *n* hypothesis and weights $\{h_i, \pi_i\}$ generated in this way increasingly approximates the posterior $P(h_t / \vec{e}_t)$ as *n* increases[3].

There are 2 key questions in the previous presentation of the Factoring Sampling algorithm: how to sample from the sum in (6)?, and how to obtain each weight factor π_i?. For the first question, assuming that one has an initial set of samples of the posterior density $P(h_t / \vec{e}_t)$ and knowledge about the dynamics of the system, it is possible to use the composition algorithm [10] to obtain a set of fair samples from the sum in (6). In the case of the second question, the weight factors π_i's can be obtained in a similar way to the Likelihood Weighting algorithm by evaluating the fitness between each sample hypotheses and the observations.

In summary, to be able to keep a sample version of the posterior density in time using the Factored Sampling algorithm, one needs an initial approximation of the posterior density, and knowledge about how evaluate the likelihood and propagation densities $P(e_t / h_t)$ and $P(h_t / h_{t-1})$. The following example shows the power of Factored Sampling to approximate an arbitrary distribution.

Figure 1a shows the initial frame of a video sequence that contains 2 yellow boxes that move freely around the image plane. Factored Sampling was used to track the box in the lower right side of the image using color information. Simulating the role of a *detector agent*, the initial box to be tracked was manually selected using the mouse. In figure 1b, the blue square inside the lower yellow box shows the area selected for tracking. Identification numbers "1" and "2" were added inside the boxes to facilitate the identification of each target by the reader, but they are not part of the video sequence used to track the targets.

[3] For a proof of the Factored Sampling method see [9].

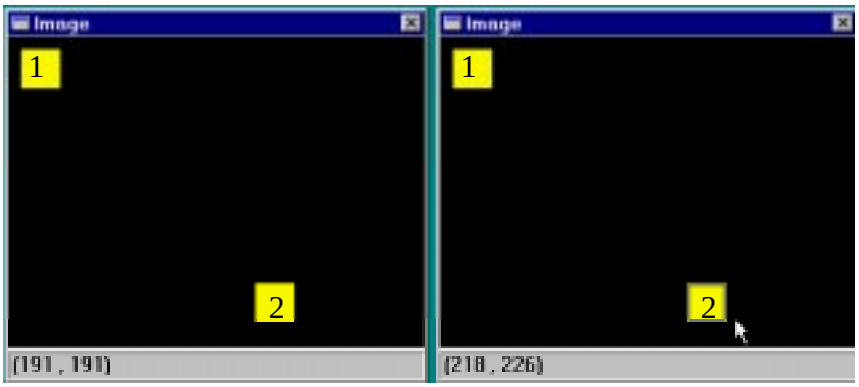


Figure 1. a) Artificial scenes with two yellow boxes. b) The intended target is manually marked with a bounding box.

Each hypothesis about the target position was given by a bounding box defined by height, width and the (x,y) coordinates of its center of mass. For this example, the height and width of each bounding box was kept fix equal to the initial area selected manually for tracking. The posterior density was at all times approximated using 500 samples. The initial approximation to the posterior density was obtained by sampling from a Gaussian distribution centered at the position of the manually selected box and with a variance of 20 pixels in both axes. Figure 2a shows an image with the initial distribution of the sample hypothesis. Figure 2b shows the Gaussian distribution of the center of mass of the hypothesis in the x-axis.

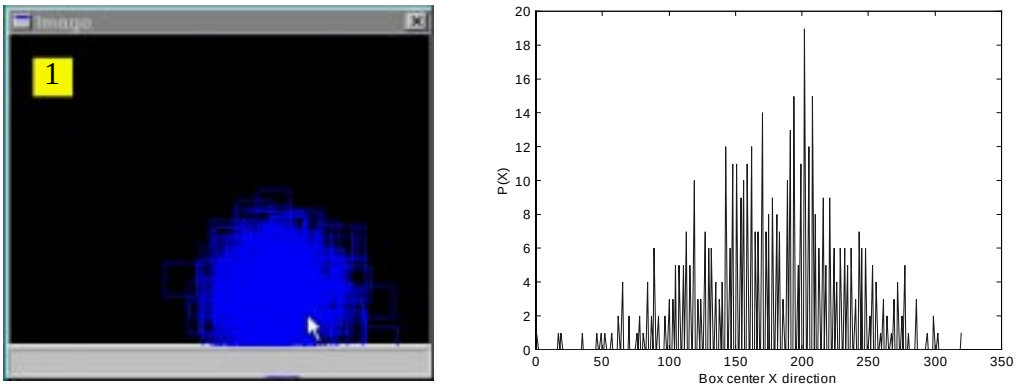


Figure 2. a) Initial distribution of the boxes. b) Distribution in the x-axis.

The tracker used as observation the hue histogram of the pixels inside each hypothesis in the sample set. The hue histogram of the initial box selected manually was kept as the desired reference. In this way the evaluation of the likelihood of each bounding box was calculated measuring the similarity between its hue histogram and the reference histogram. The metric used to evaluate similarity was a modified version of the L1 distance using a sigmoid type of function.

$$Likelihood = 1.0 - \tanh\left(2.0 * \frac{(L1\,distance - cte)}{cte}\right) \qquad (8)$$

Figure 3 shows the variation of likelihood in function of the L1 distance for a value of *cte* equal to 25% the size of the initial box.

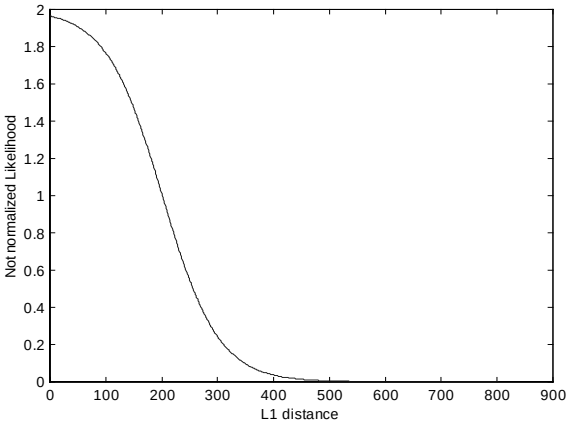
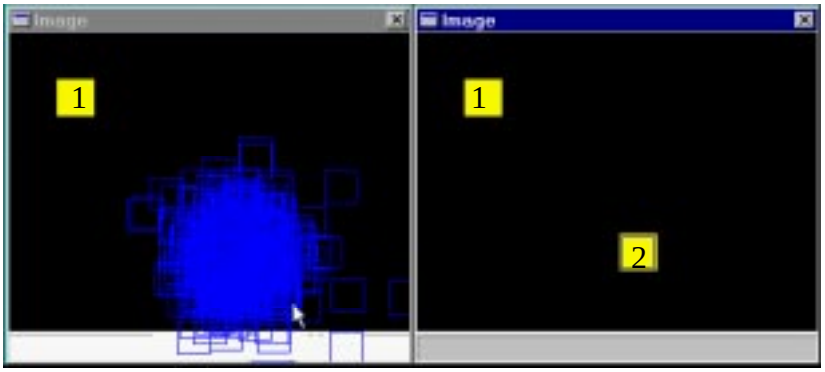
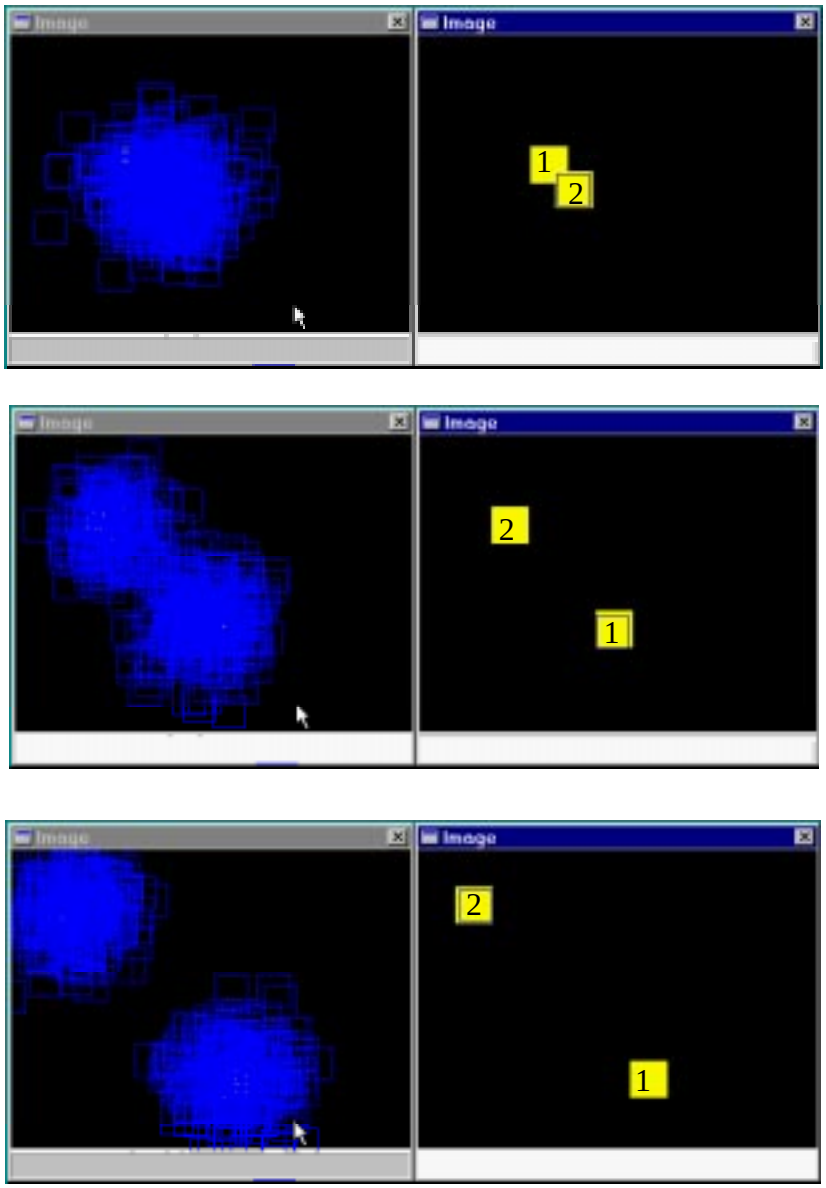


Figure 3. Variation of *likelihood in function of L1 distance for a value of cte equal to 25% the size of the original box.*

The propagation density was calculated using a stationary Gaussian model with a variance of 20 pixels in both axes. A binary search algorithm over cumulative probabilities of the current approximation of the posterior density was used to select the set of samples used for time propagation.

Figures 4-7 show the tracking at different time instants. The left figures show in blue the set of bounding boxes used to approximate the posterior density. The right figures only show the bounding box corresponding to the maximum likelihood hypothesis. The figures show the ability of Factored Sampling to dynamically approximate the posterior density. In particular, when the two identical yellow boxes overlap the tracking system becomes confuse because it does not have any way to resolve the inherent ambiguity. This ambiguity materializes in the approximation of the posterior density with a bimodal shape that is clear in figures 6 and 7. Also the maximum likelihood estimator jumps from one box to the other reflecting the high confusion in the tracker.





Figures 4-7. Left) Posterior distribution of bounding boxes for different time instants. Right) The maximum likelihood hypothesis is highlighted with a blue square inside the most probable target.

• *Bayes Nets*

Other relevant component of the framework proposed in this work is the integration of information. This integration will be performed using Bayes nets [12]. Bayes nets take advantage of causal relations among random variables to allow an efficient graphical representation of joint probability distributions (jpds). The efficiency is gained by use of causal knowledge that provides conditional independence relations between the random variables. These independence relations allow partitioning the jpd in simpler local probabilistic models.

Figure 8 shows the typical tree structure of the Bayes nets relevant to this work. Agent nodes represent the *detectors* and *specialists* able to directly measure visual properties from the incoming information. Abstraction nodes allow the integration of information

and the representation of relevant visual structures. Also, abstraction nodes allow introducing conditional independence relations among the agents. This decoupling of the information provided by the agents facilitates the construction of probabilistic models for applying Bayesian inference using equation (6).

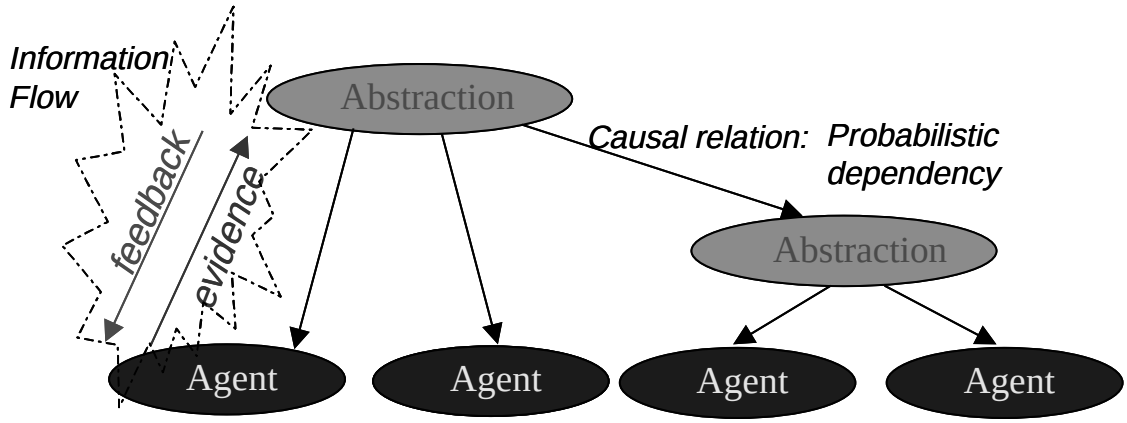


Figure 8. Example of a Bayes net structure.

One of the most important features of the net is the flow of information. Agent nodes send evidential information to high-level abstractions nodes in the form of likelihood functions represented by Gaussian densities or stochastic sampling. In the same way, abstraction nodes send top-down feedback in the form of priors to the agent nodes. These priors allow an efficient communication channel from *detectors* to *specialists* allowing the self-adaptation of the system and the generation of opportunistic behaviors on the agents.

This type of information flow is closely related to the idea of bottom-up and top-down information paths in the human visual system [22]. After the work of Marr [13] the bottom-up model has been the usual approach in machine vision. One of the main problems of a pure bottom-up model is the lack of feedback from high to low-level stages. A bottom-up model is not able to learn and use knowledge in order to reduce the heavy load of a powerful vision system. Each time, a bottom-up model needs to reprocess all the information looking for the desired cues and answers. The use of a top-down feedback from high level inference modules can guide the vision system to search for the correct visual cues and answers in the correct places, which can produce a great impact in the robustness and efficiency of such a system.

Bayes nets provide a suitable framework to implement bottom-up and top-down information channels. Each time that a *detector agent* finds a new relevant structure a new Bayes net is instantiated. The information about the new structure is sent by the *detector agent* to the abstraction node, which then sends the information as priors to a set of *specialist agents*. Each *specialist* initiates a predictive step over the feature. The *specialists* with a low self-evaluation will stop working, and the ones with a high evaluation will assume the work of keeping track of the feature.

• *Adaptation*
A difference of most traditional applications of Bayes Nets, where the structure of the nets is fix, the system intended in this research will include adaptation mechanisms that will allow a dynamic reconfiguration of the nets according to the characteristics of the incoming visual information. So far, there are 2 main adaptation mechanisms that will be included in the intended system: Agent switching and Belief sampling.

Agent switching refers to the pro-activation of *detector* and *specialist* agents. Using self-evaluation functions these agents will locally estimate the benefits of providing information to the inference engine. This evaluation will be based on robustness and efficiency considerations. An initial implementation of a self-evaluation function is discussed in section 4.

Belief sampling refers to the number of samples used by the system to keep track of the posterior density or belief function. This variable plays an important role in the computational complexity of the system because it defines the number of hypothesis to be considered by each active agent, and the fusing nodes. An initial heuristic approach to decide about this variable is discussed in section 4.

3. RELATED WORK

The idea of reducing uncertainty by combining knowledge from difference sources is by not account new. In several fields it is possible to find studies that recognize the relevance of integrating information in order to create more robust and flexible systems. Although all the abundant literature, there have been a gap between the conceptual idea and the production of working systems for real problems. Important issues such as the organization and control of the pieces of knowledge, and in special the development of mechanisms that allow the adaptation and feedback among the knowledge sources have not been tackled in depth, and they are still very much open questions. This section reviews some of the main efforts that have been appeared in the scientific literature under these lines.

• *Artificial Intelligence (AI)*
In the AI domain the blackboard model for problem solving is one of the first attempts to adaptively integrate different types of knowledge sources. Using ideas independently proposed by Newell [14] and Simmon [15], Reddy and Erman implemented the first blackboard systems as part of the HEARSAY and HEARSAY II speech understanding programs [16][17].

A blackboard model consists of 3 major components: the knowledge sources, the blackboard, and the control unit. A blackboard model divides a problem in *knowledge sources*, which are kept separate and independent. These knowledge sources interact through a *blackboard*, which is the global database that integrates the information. Finally, a *control unit* manages the opportunistic activation of the knowledge sources according to changes in the blackboard.

The blackboard conceptualization is closely related to the ideas presented in this work, but as a problem-solving scheme the blackboard model offers just a conceptual framework for formulating solutions to problems. In this sense, at least for the 2 applications presented here, the work proposed in this research aims to extent the blackboard conceptualization to a computational specification or working system, providing specific mechanisms to perform probabilistic inference and adaptive integration of visual information.

• *Machine Learning*

In the machine learning literature there are been related work in the context of ensembles of classifiers. An ensemble of classifiers is a set of classifiers whose individual decisions are combined to classify new examples [18]. Each classifier can be considered as a different source of knowledge. Adaptation mechanisms are included in the policy used to combine the outputs of the individual classifiers. These kinds of techniques are currently receiving broad attention in the machine learning literature due to the capacity of the ensemble to improve performance over the individual classifiers that make them up. There have been several algorithms proposed to implement the ensemble of classifiers; among the more relevant are *Mixture of Experts* [19] and *AdaBoost* [20].

The work presented in this proposal differs in many ways with respect to the current algorithms used to build ensemble of classifiers. One of the main differences resides in the adaptation mechanisms. An ensemble of classifiers is an eager learner in the sense that the training is performed off-line and during operation each classifier acts as blind data driven box. In contrast, one of the main features of the work proposed here is the on-line interaction or feedback between the knowledge sources.

• *Computer vision*

In the area of computer vision although most of the work has been concentrated in the development of algorithms to extract knowledge from single visual cues, such as edges or binocular disparities, there have been also several attempts to integrate information from different visual modules.

Since at least the work of Marr [13] it have been widely accepted that different visual cues should be computed in separate modules, but there has been lot of controversy about how these modules should interact to create a unified visual representation. Marr suggests that the information from different visual modules is integrated to obtain a complete, labeled 3D description of the world (2 ½-D sketch). According to Marr, in this process there is not interaction between the different visual modules, and perceptual vision is just a bottom-up data driven process. Recently, several researchers have criticized this idea and proposed an alternative model [21][22], which considers perception as a distributed collection of task-specific, task-driven visual routines with strong feedback among the visual modules.

Besides all this conceptual debate, the strange fact is that although the constant acknowledge in the computer vision literature about the importance of integrating visual

information, there have been not many working systems that exploit these ideas. There are several reasons to argue about this situation. One of the main reasons is that as a new field rather than looking for the complex problem of integration, the focus of the research has been concentrated in solving more elemental issues, such as increasing the robustness of individual visual modules. Also there has been important hardware limitations because of the intensive computing power needed to run several visual modules in parallel. Other possible reason is a lack of a global perspective to envision vision as an interdisciplinary field including elements from areas such as information theory, statistics, and artificial intelligence, which provide more solid theories to implement the integration and adaptation steps.

Although the previous argumentation, still it is possible to find some interesting works that share some of the ideas presented in this research proposal. Krotkov and Bajcsy [23] used a Kalman Filter approach to combine stereo, depth from focus, and vergence in order to obtain more robust 3-D data. Brautigam [24] used a voting scheme among several visual modules to detect planar surfaces. Nordlund and Eklundh [25] described a system that integrates motion and binocular disparities to achieve real time figure ground segmentation. In the context of object recognition and image understanding, Drapper [26] investigated the use of learning strategies to determine recognition policies using different visual modules. Darrel et al. [27] presented a tracking system that combines information from stereo, color, and face pattern matching. Isard and Blake [11] proposed a probabilistic approach to target tracking under a Bayesian framework. Sherrah and Gong [28] proposed the use of covariance estimation to track pose and face position fusing skin color information and pose similarity measures. Rasmussen and Hager [29] proposed an adaptive visual system that integrates information from several visual cues using the Joint Probabilistic Data Association Filter (JPDAF).

Although most of these works have shown the gain in robustness of combining several visual modules, most of them have limited the scope to static scenes and lab implementations. Even more important, most of these works have not considered in their systems topics such as adaptation and uncertainty. The works by Isard and Blake, and Rasmussen and Hager are notable exceptions, both will be discussed in more detail later in this document.

- *Cognitive Psychology*

It is worth to mention that one of the main motivations for the integration of visual cues comes from studies in the human visual system. The theory of visual specialization has become widely accepted by the cognitive psychology community as a partial explanation about how the brain achieves visual perception [22]. According to this theory different attributes of the visual scene such as form, color, motion, and depth are processed in parallel in different areas of the cerebral visual cortex. Unfortunately, as in the case of computer vision, so far the studies on the human visual system have fail to explain how the knowledge from the different visual modules is put together to give us our unitary experience of the visual world.

• *Computer Vision and Target Tracking*

There have been several attempts to build visual systems for the case of dynamic target tracking. Toyama and Hager [30] presented a target tracking architecture based on an incremental focus of attention. The system consists of different trackers organized in a top-down fashion where trackers in the top layers provide higher tracking precision but poor target reacquisition capabilities and less robustness again changes in the tracking conditions. After each tracking cycle the system can adaptively move up and down the layers depending of the success of the tracking during the previous cycle. If the current tracker fails to detect a target, the system move down passing control to a tracking algorithm with better target reacquisition capabilities. One of the problems with this system is that the switching between trackers is fix without mechanisms to swap layers in and out. Also, there is not estimation of ambiguity or a mechanism to integrate information from different sources. Besides that the work did not mention how each tracker evaluates its performance. The authors claim good results for tracking several objects in indoor scenes. Unfortunally the analysis of the results is performed only in qualitative terms.

Darrel et al. [27] presented a tracking system that combines information from stereo, color, and face pattern matching. Stereo is used to segment out the silhouette of possible people. After this, a color detector specially tuned to detect skin color analyzes the candidate silhouettes searching for skin regions. This analysis provides the position of faces, which are then used by a face pattern-matching algorithm. Although most of the processing is achieved in a serial reduction, at the end the system integrates the information from color, face pattern matching, and height of the silhouette to establish target correspondence. The system was tested on real indoor images with encouraging results. According to this study the integration of information considerably improved the system performance. In contrast to the research proposed here, in this work there is not adaptation or interaction between the different vision modules. Also the decision rules used to establish target correspondence are based on maximum likelihood estimators and heuristic thresholds without considering the complete belief or posterior distribution.

Sherrah and Gong [28] proposed the use of covariance estimation to track pose and face position fusing skin color information and pose similarity measures. The tracking is based on the Condensation algorithm. The covariance of the modules is estimated from training examples, and it is used to estimate the state propagation density. Correlation between face and head positions is used to model the state-conditional density function. The experimental results show that the system is able to closely track the head pose. Also, the authors highlight the fact that without the use of data fusion the pose tracker does not track reliable at all. In this work there is not adaptation or feedback between the different vision modules.

Inspired by ideas from the statistic community, Isard and Blake proposed the Condensation algorithm [11]. The Condensation algorithm uses an extended version of Factored Sampling to keep in time a population of hypothesis about the target states and their posterior probabilities. Using learned models about the target dynamics the posterior probabilities are propagated in time using a dynamic version of Bayes rule. Isard and

Blake used the Condensation algorithm to track deformable contours in highly cluttered environments with great success. Recently, using the statistical technique of important sampling, they have extended the Condensation algorithm to consider color information as an auxiliary source of knowledge to sample from the posterior distribution [31]. Although the work of Isard And Blake only indirectly touches the ideas of integration and adaptation, their probabilistic representation is one of the main attempts in the computer vision community to explicitly represent ambiguity in the form of a belief function or posterior probability.

One of the main problems in target tracking is the data association, i.e., how to distinguish which measurements come from a specific target. The target tracking community has been studying this problem for a long time given origin to several algorithms such as the track-splitting algorithm, the joint likelihood filter, the multiple hypothesis filter, and the joint probabilistic data association filter. The naïve solution to the data association problem is the nearest neighbors approach, however in the case of dynamic target tracking it is possible to achieve more robust associations by postponing the decision process until future measurements can clarify current ambiguities. This is one of the main robustness of the Isard and Blake's Condensation algorithm. Keeping in time a sample version of the posterior density makes possible to track alternative hypothesis hoping that the new incoming information will resolve possible ambiguities. The research proposed here aims to go one step forward. Instead of passively waiting that the incoming data will resolve ambiguities, the idea of this work is to actively search for new information integrating and adaptively switching in and out visual algorithms.

Using a modified version of the Joint Probabilistic Data Association Filter (JPDAF), Rasmussen and Hager [29] presented an adaptive system that integrates visual information to perform tracking tasks. The JPDAF is an extension of the Kalman filter to deal with the data association problem in a Bayesian framework. The JPDAF modifies the innovation vector in the Kalman Filter update equation with a combined weighted innovation term. This combined innovation term is obtained by weighting each measurement with its probability to belong to a specific target. Rasmussen and Hager extended the JPDAF in order to deal with information from several visual cues and mutual constraints between targets. Interesting is the fact that using information from the strength of the associations between targets and measurements, they are able to adaptively switch visual cues in and out. Rasmussen and Hager have tested this system for tracking multi-part objects with promising preliminary results.

One of the main merits of the work of Rasmussen and Hager is that it includes ideas such as adaptation, ambiguity, and integration in a working system. As far as I know, this is the only working system with such features. Although their system has shown promising results, the JPDAF has several limitations as a framework to perform these kinds of tasks. One of the main limitations is the averaging or expected value calculation used by the JPDAF to obtain the combined innovation factor. This averaging is a consequence of the normality assumption used by the Kalman Filter, but it is not valid in a general case. The risk of using averaging is particularly significant for the case of conflicting measurements because the tracker can end up tracking an average position where there is not target at

all. Another limitation of the JPDAF is that it just provides weak mechanisms to characterize false measurements and feasible hypothesis. This is especially problematic because of the combinatorial explosion in the number of data associations. On other hand, the framework described by Rasmussen and Hager does not include a mechanism to initialize the trackers.

The work of Rasmussen and Hager is closely related to the research intended in this proposal. The main difference resides in the mathematical machinery used to solve the problem. While Rasmussen and Hager use a modified version of the Kalman filter, the work presented here uses Bayesian Nets in conjunction with Factored Sampling to keep an approximation of the posterior density. I believe that incorporating measurement of reliabilities based on the posterior density rather on heuristics will produce a more robust system.

• *Obstacle Detection*
In the mobile robotics domain there is an extensive list of works about the use of vision to solve the obstacle detection problem. Most of the working systems can be classified into two categories: depth estimation or use of low-level visual cues.

There have been several attempts to detect obstacles using a geometrical reconstruction of the depth structure of the environment. Although there have been some attempts using visual cues such as focusing [32] and depth from motion [41], binocular stereo has been the favorite method used in most of the systems [33] [34]. At a first glance the use of stereo vision seems very appealing. A robust system able to generate correct binocular matches for every point in a scene could be the key to solve not only the robot navigation, but also other recognition problems. Unfortunately, so far the state of the art shows that this robust system is not possible. Textureless areas, occlusion, repeated patterns seem to be ill-posed problems. Also calibration problems and the high computational requirements make this approach even more difficult. Ratler and Nomad at CMU, and Robby and HMMWV at JPL are examples of mobiles robots that use a stereovision system to detect obstacles.

The limitations of stereovision to produce an accurate 3D description of the environment have produced a change in the scope of these systems. In this way, ideas such as evidence grid and pyramidal correspondence have appeared in the literature [35]. In [36] under the title “Why stereo vision if not only always about 3D reconstruction”, W. Grimson presented an interesting alternative approach where stereo is used to determine figure/ground separation instead of 3D reconstruction. Here, instead of finding absolute depth estimation, the role of stereo is to find cluster of neighboring points that have similar depth. This is a good example of a new tendency to use qualitative rather than metric information. In general qualitative information is easier to obtain, and it is more robust to noise than metric information.

After the work of I. Horswill with the robot Polly [37], there has been an increased interest in exploiting low-level visual cues to detect obstacles. These systems use task and environment constraints to simplify the detection of obstacles using just low-level visual

cues such particular colors, textures, shapes, and so on. For example, [38] describes a mobile robot that is able to navigate on particular indoor environments detecting the texture properties of the floor.

The main characteristic of these systems is their simplicity. Typically, these vision systems are able to run at several cycles per second. This high efficiency is achieved due to a fine-tuning between the perception capabilities and the task/environment constraints. However this simplicity can produce several failure modes when the assumptions are slightly violated.

One of the main limitations with the use of low-level visual cues is that so far there is not an automatic procedure that allows selecting appropriate visual cues. Also, there is a lack of a clever way to fuse the information from different visual cues. Most of the systems base the fusion on simple average or voting schemes.

4. CURRENT PROGRESS

A preliminary version of the system proposed in this work has been implemented using color and stereo visual cues. These cues were used to implement two *detector* and two *specialist* agents[4]. The color *specialist* is based on hue and the stereo *specialist* on depth continuity. These *specialists* evaluate local likelihood functions in a similar way that the color *specialist* described in section 2.3.

The agents have been implemented using CyberAries (Autonomous Reconnaissance and Intelligent Exploration System) [39], a multi-threaded and distributed agent architecture that greatly simplifies the job of developing ensembles of cooperating agents. Each agent runs as a separate thread in a local or remote machine using standard sockets to communicate to the rest of the system.

The preliminary system was evaluated for the case of single person tracking and obstacle detection. Bounding boxes were used to describe the state of a tracked target or a detected obstacle. These bounding boxes were modeled by their center position, width, and height using a multivariate Gaussian density of known mean and diagonal covariance matrix. Factored Sampling and Bayesian fusion using equation (6) were used to maintain an approximation of the posterior pdf using 1000 samples.

- *Single Person Tracking*

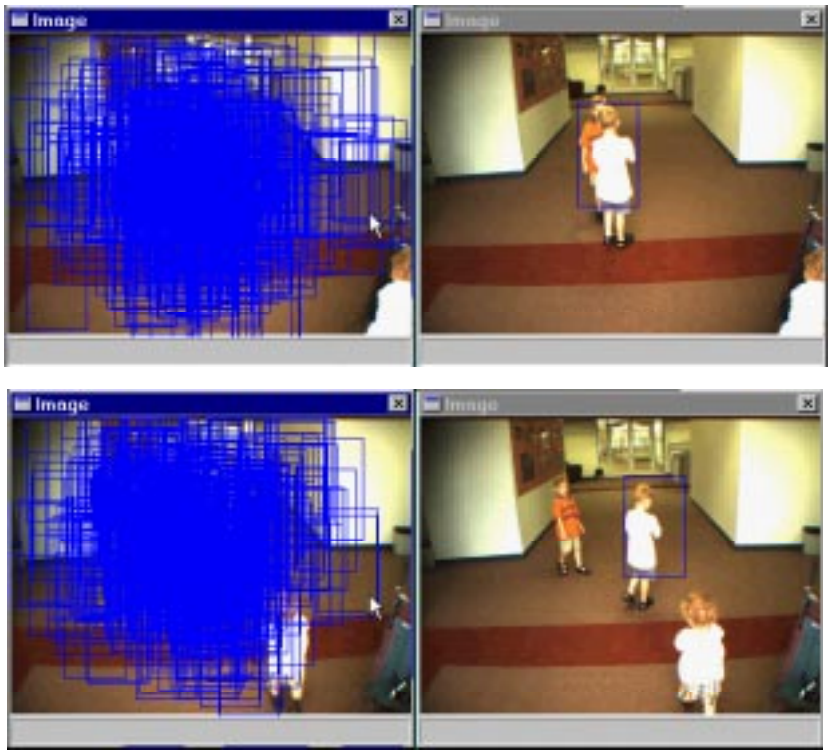
In order to evaluate the benefits of the framework presented in this research proposal, three different schemes were used to track a single person in the same video sequence. Figure 9 shows the intended target enclosed by a bounding box.

[4] See [40] for a description of the segmentation algorithm used to build the *detectors*.

Figures 9. Initial detection of an intended target by a detector agent.

Scheme 1: Tracking using stereo

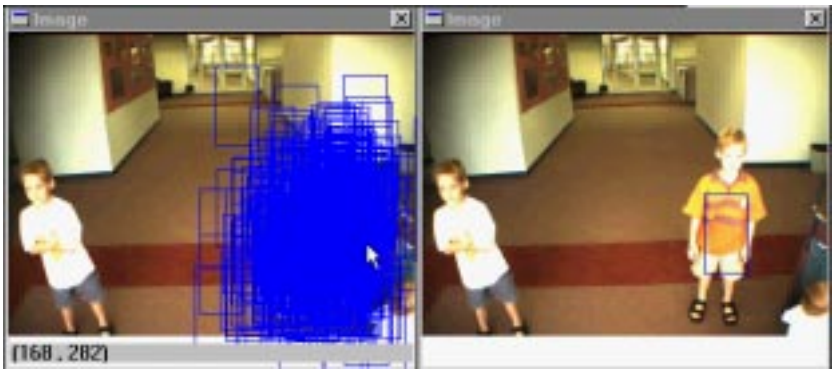
The first scheme uses a tracker based only on 3D positions provided by a stereovision system. This tracker set up a baseline to compare the benefits of adaptively integrating different visual modules. Figures 10-13 show the performance of this tracker for different time instants. The left figures show in blue the set of bounding boxes used to approximate the posterior density. The right figures show the bounding box of the maximum likelihood hypothesis. Notice how the tracker becomes confuse when the tracked person walks close to another person. In special, notice how the maximum likelihood hypothesis becomes erroneous.

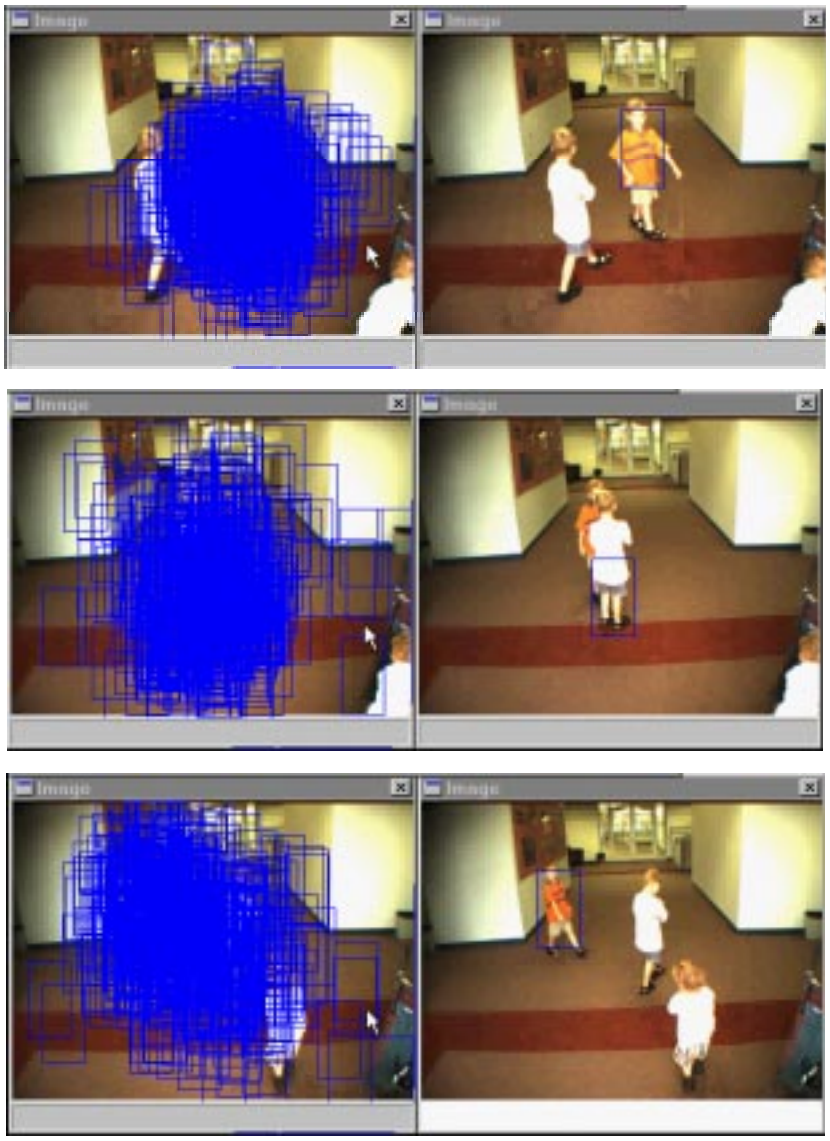


Figures 10-13. Left) Distribution of bounding boxes for different time instants. Right) Bounding box with the maximum likelihood.

Scheme 2: Tracking using stereo and color information.

The second scheme uses a tracker based on stereo and color information. Figure 14-17 show the performance of this tracker for different time instants. Notice how the approximation of the pdf given by the blue boxes is more compact than the previous scheme. This more compact or unimodal distribution shows the reduction in ambiguity obtained by adding the color information. Notice also how the system is able to re-acquire the target after a momentary occlusion. For this example the system was able to run in real time with an average processing time of 3.2 hz.





Figures 14-17. Left) Distribution of bounding boxes for different time instants. Right) Bounding box with the maximum likelihood.

Scheme 3: Adaptive tracking using stereo and color information.

The third scheme uses an adaptive tracker based on stereo and color information. The adaptation is based on local self-evaluation functions on each agent. After each tracking cycle each agent compare its local likelihood function with the likelihood obtained by the fusing nodes. This comparison is based on a discrete version of the Kullback-Leibler divergence defined by:

$$D(f,g) = \sum_i f(i) * \log \frac{f(i)}{g(i)} \qquad \text{with } f(i) = \text{normalized likelihood at sample } i$$

Also each agent keeps a normalized processing time, obtained by normalizing the local average processing time with respect to the processing time of the other active agents. These normalized processing time is multiplied by the local *D(f , g)* to obtain a local

performance score. This performance score is used by each agent to start or stop processing according to the degree of ambiguity measured by the fusing node.

After each cycle the fusing node calculates the entropy E of the approximation to the posterior density.

$$E = \sum_i f(i) * \log_2 f(i)$$

If the entropy is lower than a threshold the fusing node sends a message to the agent with the lower performance score to stop processing. In the same way if the entropy is greater than a threshold the fusing agent sends a message to start any non-active agent. Also, in the case that the entropy is lower that a certain threshold and there is only one agent active, the adaptive system decreases the number of samples used to estimate the posterior probability by 10%.

Compared to the results obtained using the scheme 2, the adaptive system was able to speed up the processing time by a factor of 2.8 without major difference with respect to the tracking performance. Figure 18 shows the adaptive configuration of the Bayes net. After frame 15 the system decided to operate only with the color visual agent.

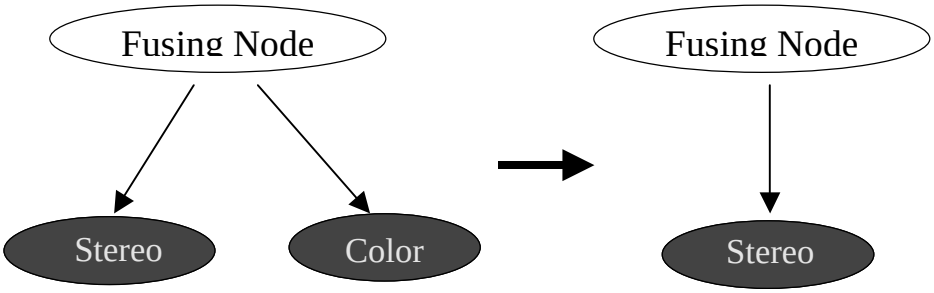


Figure 18. Adaptive evolution of the Bayes net.

• *Obstacle Detection*

Figure 19 shows an example of the performance of the system for the detection of obstacles using stereo and color information. The upper images show the detection based only on stereo for different time instants in the video sequence. From this figure it is clear that the effect of noise makes not possible a robust tracking of the features using just the stereo agent. The lower images show the combined tracking based on the color and the stereo agents. Combining both cues the system was able to keep track of all the structures during the complete video sequence consisting of 40 video frames.

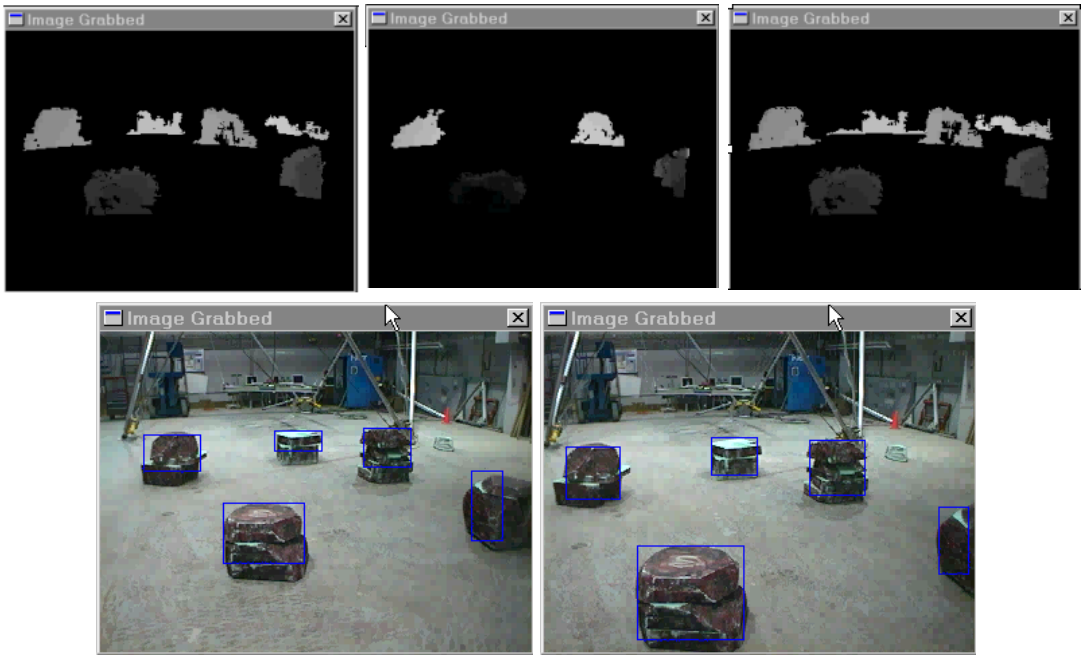


Figure 19. Upper images show the structures detected by the stereo agent at some points during the robot motion. Lower images show the detection on the obstacles for the initial and final frame in the video sequence using information from the color and the stereo agents.

5. CONTRIBUTIONS

- To develop a new framework to perform visual tasks through the creation of an adaptive visual system able to selectively combine a wide variety of visual information.
- A synergistic combination of elements from computer vision, intelligent agents technology, probabilistic reasoning, and information theory for the creation of a *flexible*, *robust* and *efficient* vision system.
- To develop innovative metrics to express under probabilistic terms the fitness between visual hypothesis and observations
- To study innovative adaptation criterions and their optimality conditions
- To evaluate the ideas presented in this proposal in a working system to perform target tracking and obstacle detection by a mobile robot
- To recommend improvements and new lines of investigation according to the results obtained with this research.

6. FUTURE SCHEDULE

Fall 2000

- Incorporate agents based on motion and texture
- Study new adaptation criterions and their optimality
- Develop an adaptation mechanism to control the number of samples

Spring 2001

- Develop a reinforcement learning algorithm to learn optimal switching strategies
- Evaluate alternative propagation models
- Initiate the evaluation period

Summer 2001

- Complete evaluation period
- Compare results with alternative methods

Fall 2001

- Complete written thesis
- Defend

7. REFERENCES

[1] A. Soto and I. Nourbakhsh. *"A Scenario for Planning Visual Navigation of a Mobile Robot"*, AAAI 1998 Fall Symposium Series, October 23-25, Florida, 1998.

[2] I. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto. *"An Affective Mobile Educator with a Full-time Job"*, Artificial Intelligence, Vol. 114, No. 1 - 2, October, 1999, pp. 95 - 124.

[3] C. Diehl, M. Saptharishi, J. Hampshire II, and P. Khosla. *"Collaborative Surveillance Using Both Fixed and Mobile Unattended Ground Sensor Platforms"*, forthcoming in Proceedings of SPIE, Vol. 3693, AeroSense, Orlando, Fl., April, 1999.

[4] A. Lipton, H. Fujiyoshi, and R. Patil. *"Moving Target Classification and Tracking from Real-time Video"*, Proc. of the 1998 DARPA Image Understanding Workshop (IUW'98), November 1998.

[5] D. Pomerleau. *"Neural Network Perception for Mobile Robot Guidance"*, Boston: Kluwer Academic Publishers, 1993.

[6] N. Jennings and M. Wooldridge. *"Applying Agent Technology"*. Journal of Applied Artificial Intelligence special issue on Intelligent Agents and Multi-Agent Systems, 1995.

[7] Peter Maybeck. *"Stochastic Models, Estimation, and Control"*, Academic Press, Vol. 1 – 1979.

[8] R. Shachter and M. Peot. *"Simulation Approaches to General Probabilistic Inference on Belief Networks"*. In proc. 5th conference on uncertainty in artificial intelligence, 1989.

[9] U. Grenander, Y. Chow, and D. Keenan. *"HANDS. A Pattern Theoretical Study of*

Biological Shapes", Springer-Verlag, New York, 1991.

[10] M. Tanner. *"Tools for Statistical Inference"*, Springer series in statistics, 3th edition.

[11] M. Isard and A. Blake. *"Visual Tracking by Stochastic Propagation of Conditional Density"*, Proceedings of 4th European conf. on computer vision, 343-356, Cambridge, England, 1996.

[12] J. Pearl. *"Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference"*, The Morgan Kaufmann Series in Representation and Reasoning, 1991.

[13] D. Marr. *"Vision"*, San Francisco: freeman, 1982.

[14] A. Newell. *"Some problems of basic organization in problem-solving programs. Conference of self-organizing systems"*, Conference of self-organizing systems, Washington D.C.: Spartan books, 393-42, 1962.

[15] H. Simmon. *"Scientific Discovery and the Psychology of the Problem Solving"*. In models of discovery, Boston, Mass: D. Reidel Publishing company, 1977.

[16] R. Reddy, D. Erman, and N. Richard. *"A model and a system for machine recognition of speech"*, IEEE transaction on audio and electroacustic AU-21:229-238, 1973.

[17] D. Erman, F. Hayes-Roth, V. Lesser, and R. Reddy. *"The HEARSAY-II speech understanding system: integrating knowledge to resolve uncertainty"*. ACM computing survey 12:213-253, 1980.

[18] T. Dietterich. *"Machine Learning Research: Four Current Directions"*, AI magazine, 18 (4), 97-136, 1997.

[19] S. Waterhouse. *"Classification and Regression using Mixtures of Experts"*, PhD. thesis, Cambridge University, October 1997.

[20] S. Freund and R. Schapire. *"A Decision Theoretic Generalization of On-Line Learning and an Application of Boosting "*. Proc. of the 2th European conference on computational learning theory, pp. 23-37, Springer-Verlag, 1995.

[21] S. Ullman. *"Visual Routines"*, Cognition, 18, 1984.

[22] S. Zeki. *"A Vision of the Brain"*, Oxford, Blackwell scientific publications, 1993.

[23] E. Krotkov and R. Bajcsy. *"Active Vision for Reliable Ranging: Cooperating Focus, Stereo, and Vergence"*, Intl. Journal of Computer Vision, vol. 11, no. 2, October 1993, pp. 187-203.

[24] C. Brautigan. *"A Model Free Voting Approach to Cue Integration"*, PhD. Thesis, Stockhoms Universitet, August 1998.

[25] P. Nordlund and J.-O. Eklundh. *"Real-time maintenance of figure-ground segmentaion"*, in Proc. 1st Int. conference on computer vision systems, vol. 1542 of lecture notes in computer science, pp. 115--134, Springer Verlag, Berlin, Jan. 1999.

[26] B. Draper and A. Hanson. *"An Example of Learning in Knowledge-directed Vision"*, Scandinavian conference on image analysis, Aalborg, DK., August 1991. pp. 189-201.

[27] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. *"Integrated Person Tracking Using Stereo, Color, and Pattern Detection"*, Proc. of the conference on computer vision and pattern recognition, pp. 601-609, Santa Barbara, June, 1998.

[28] J. Sharra and S. Gong. *"Fusion of Perceptual Cues using Covariance Estimation"*, Proceedings of BMVC'99, 13-16 September 1999, Nottingham, England.

[29] C. Rasmussen and G. Hager. *"Joint Probabilistic Techniques for Tracking Multi-Part Objects"*, Proc. of the conference on computer vision and pattern recognition, Santa Barbara, June, 1998.

[30] K. Toyoma and G. Hager. *"Incremental Focus of Attention for Robust Visual Tracking"*. Proc. of the conference on computer vision and pattern recognition, 1996.

[31] M. Isard and A. Blake. *"ICONDENSATION: Unifying Low-Level and High-Level Tracking in a Stochastic Framework"*, Proceedings of 5th European conf. on Computer Vision, 893-908, Cambridge, England, 1998.

[32] I. Nourbakhsh. *"A sighted Robot: Can we Ever Build a Robot that Really Doesn't Hit(or fall into Obstacles?"*, The robotics practitioner, spring 1996, pp.11-14.

[33] E. Krotkov et al., *"Stereo Perception and Dead Reckoning for a Prototype Lunar Rover"*, Autonomous Robot 2(4), pp. 313-331, December 1995.

[34] L. Matthies. *"Stereo Vision for Planetary Rovers: Stochastic Modeling to Near Real-Time Implementation"*, International Journal of Computer Vision, 8(1):71-91, July 1992.

[35] M. Martin, *"Breaking out of the black box, a new approach to robot perception"*. Thesis proposal, Robotics Institute, Carnegie Mellon University, 1998.

[36] W. Grimson, *"Why stereo vision is not always about 3D reconstruction"*. MIT AI laboratory memo no. 1435, July 1993.

[37] I. Horswill, *"Polly: a Vision Based Artificial Agent"*, Proceedings of the 11th national conference on artificial intelligence (AAAI-93), July 11-15, Washigton DC, 1993.

[38] L. Lorigo. *"Visually-Guided Obstacle Avoidance in Unstructured Environments"*. MIT AI laboratory master thesis, February 1996.

[39] C. Diehl, M. Saptharishi, J. Hampshire II, and P. Khosla. *"Collaborative Surveillance Using Both Fixed and Mobile Unattended Ground Sensor Platforms"*, forthcoming in Proceedings of SPIE, Vol. 3693, AeroSense, Orlando, Fl., April, 1999.

[40] A. Soto, M. Saptharishi, J. Dolan, A. Trebi-Ollennu, and P. Khosla, *"CyberATVs: Dynamic and Distributed Reconnaissance and Surveillance Using All Terrain UGVs"*. Proceedings of the international conference on field and service robotics, August 29-31, 1999.