Active Recognition: Using Uncertainty to Reduce Ambiguity

Francesco G. Callari and Frank P. Ferrie

TR-CIM-95–11 Sept. 26, 1995
Revised Jan. 10, 1996

Centre for Intelligent Machines
McGill University
Montréal, Québec, Canada

Postal Address: 3480 University Street, Montréal, Québec, Canada H3A 2A7
Telephone: (514) 398-6319 Telex: 05 268510 FAX: (514) 398-7348
Email: cim@cim.mcgill.ca

Active Recognition: Using Uncertainty to Reduce Ambiguity

Francesco G. Callari and Frank P. Ferrie

Abstract

Ambiguity in scene information, due to noisy measurements and uncertain object models, can be quantified and actively used by an autonomous agent to efficiently gather new data and improve its information about the environment. In this work an information-based utility measure is used to derive from a learned classification of shape models an efficient data collection strategy, specifically aimed at increasing classification confidence when recognizing uncertain shapes. Promising experimental results are reported.

Résumé

L'ambiguïté de la reconnaissance d'objets, due aux mesures bruitées et à l'incertitude des modeèles des objets, peut être quantifiée et être utilisée d'une manière active par un robot autonome pour acquérir efficacement de nouvelles données et améliorer sa connaissance de l'environnement. Dans cet article, une mesure d'utilité des données, basée sur la théorie de l'information, est utilisée pour dériver à partir d'une classification acquise des modèles de formes une stratégie efficace de cueillette de données, spécifiquement orientée vers une amélioration du degré de confiance de la classifiction lors de la reconnaissance de formes incertaines. Des résultats de simulation prometteurs sont présentés et discutés.

1. Introduction

A common paradigm underlying several current approaches to model-based active object recognition can be stated in terms of the following algorithm: (a) gather sensory information, (b) build or refine a model of the environment, (c) attempt a recognition, (d) plan and take further measurements, using "context" (a-priori) knowledge and/or the model itself, whenever the tentative recognition appears to be too ambiguous. The recovery of models from the data can rapidly become a highly demanding task when high bandwidth sensory information is used, as is the case in 3D vision-based navigation [9]. This is particularly true for vision-based systems that demand "rich" models, i.e. general models with high descriptive capability (Figure 1), usually because the agent interacts with poorly structured environments, upon which few assumptions can be made a-priori. In this case the problem of gathering, registering (view-coupling), and generally managing large quantities of data is paired with the computationally expensive problem of explaining them by nontrivial models, usually by some form of optimal fit.

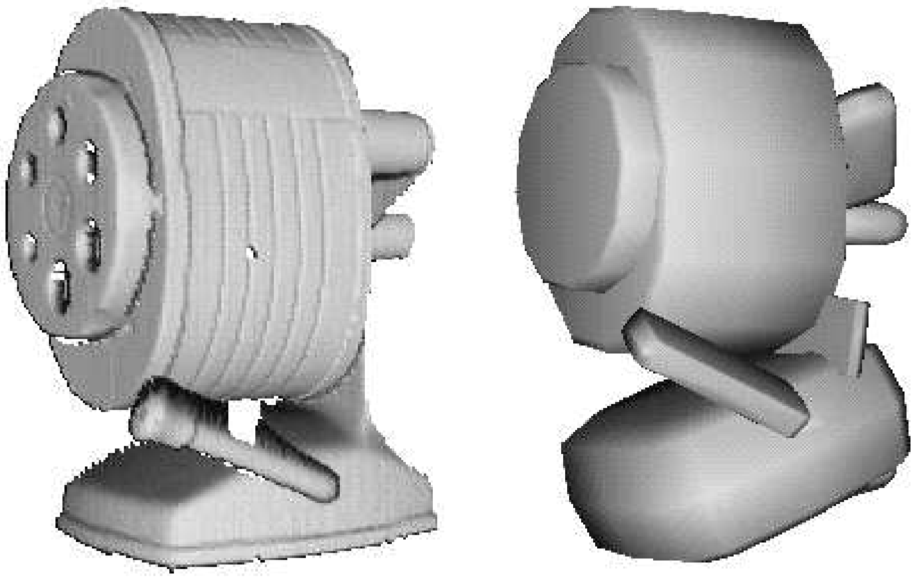


FIGURE 1. Left: a shaded image of a pencil sharpener rendered from measurements obtained with a laser rangefinder. Right: a model of the sharpener computed from the rangefinder data using superellipsoid primitives to represent the parts of the sharpener.

A further complication is added by the presence of uncertainty in the models, due to noise in the data and/or to limits in the descriptive power of the models themselves which, after all, are only approximations of a usually more complex reality. However a proper treatment of this uncertainty is *essential* in many cases, particularly when the cost of making a wrong decision with the current model is large, compared to the one involved in delaying the decision to gather further data and refine the model. A typical example might be the interaction of a robot with dangerous, potentially harmful objects like containers of toxic substances in controlled environments. At

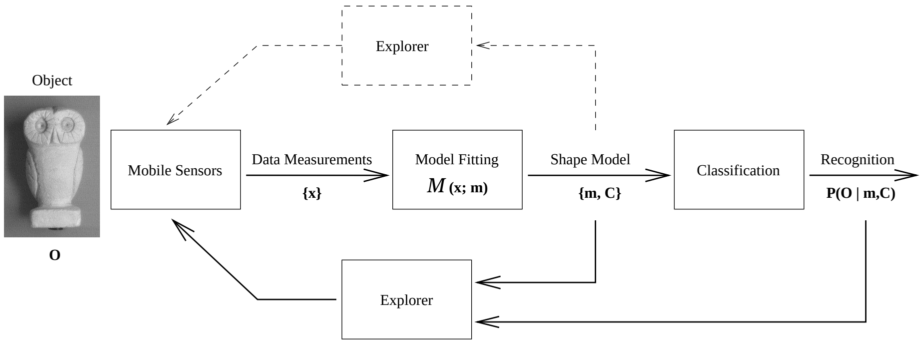


FIGURE 2. A scheme for the recognition-driven paradigm discussed in the paper. The proposed exploration loop is indicated by the solid lines; the loop in dashed lines depicts the autonomous exploration technique used by Whaite and Ferrie to model shapes.

first glance this idea of model refinement through further data collection might seem to introduce a chicken-and-egg problem: the agent should decide where to go next based on a model, but it needs to go somewhere to collect data, so as to build a better model before deciding. This apparent circularity can be resolved by observing that the current model can be used to make physical predictions about the environment. Furthermore, model uncertainty, if properly quantified, can be propagated to the prediction themselves, thus augmenting them with predictive error bars. Intuition and sound statistical arguments (the theory of optimal experiments [4]) then suggest that the best places to collect data from in order to refine the model are those where the error bars are large, since these are the data-gathering points that are most likely to provide useful information. What is a "large" error bar is a nontrivial issue, especially in the case of nonlinear models, as will be later discussed.

In this paper we deal with the above issues for the case in which the sensor actions to be planned are aimed at recognizing objects (e.g. landmarks, tools) from geometrical models of them. This active recognition problem can then be subdivided in the following subproblems:

- Gather data, merge (register) them with previous ones, fit them with shapes of given mathematical form with unknown parameters, specify the model uncertainty by quantifying the uncertainty in the parameters themselves.[1]
- Attempt object recognition, and validate the adequacy of the current uncertain shape model in terms of confidence in the recognition itself.
- Disambiguate inadequate "recognitions" by efficiently collecting new data to reduce recognition ambiguity.

Although active object recognition, in its various forms, has traditionally attracted considerable attention, the above problems can still be regarded as open. Several schemes in the literature [10, 13] pose the recognition problem as one of matching

[1] Of course each one of these steps may well be the subject of whole shelves of literature. We group them together simply because we do have working solutions for them [15, 16], hence we tend to regard them as being dealt with by a subsystem within the architecture subject of the present work.

recovered volumetric models with those of a database, using distance measures like dot product [10] or Mahalanobis distance. However the inherent uncertainty in the recovered models is rarely either estimated or accounted for in a principled way, save for *ad hoc* rules to restrict the variability of match candidates, whereas the very presence of noise in the measurements and recovery errors should lead to a probabilistic treatment. A few works have actually cast the object recognition problem in a probabilistic perspective [2, 13], yet usually pursuing an "identification" approach, aimed at producing definite statements about model matching versus a database of objects, usually via MAP estimation weighted by statistical risk. While this approach is the right (and only) choice when no further data collection is possible, it is often the case that delaying a decision and trying to reduce the uncertainty by gathering new data is predictably less expensive than attempting a decision in the current state of ignorance. Any sensible risk measure, however task-dependent, must in fact decrease with the amount of relevant information gathered.

The recognition-driven planning paradigm we refer to is illustrated in Figure 2. Along the forward path, from object to class estimates, data $\{\mathbf{x}\}$, measured by sensors on object $\mathbf{O}_i$, are fitted to a parametric shape model $\mathcal{M}(\mathbf{x};\mathbf{m})$, whose parameters $\mathbf{m}$ are recovered with uncertainty $\mathbf{C}$.[2] A classifier eventually produces a "recognition" in terms of conditional probability estimates $P(\mathbf{O}_i|\mathbf{m})$, over the possible objects given the model. The backward path (shown in solid lines in Figure 2) indicates the goal of the present work, namely to derive from the uncertainty in the object model and from the ambiguity of the classification an exploration strategy aimed at actively and efficiently collecting data in order to reduce as best as possible the ambiguity itself.

Along the same line of thought Whaite and Ferrie have proposed a technique of autonomous exploration [15, 16, 17] for model-based shape reconstruction from laser rangefinder images. Their autonomous explorer (schematically depicted by the backward path in dashed lines in Figure 2) actively seeks viewpoints from where it can observe parts of the object surface with large predicted variance from the model, thus improving the model estimate until the object surface can be predicted as reliably as it can be measured. From a theoretical point of view, this strategy is provably optimal since it maximizes the information gain in the model parameters across successive data-gathering steps [6], thus being equivalent to a 'Q-optimal' experiment design [4].

An efficient data collecting strategy aimed at confident object recognition, however, is arguably different from the above even if it's guided by the same optimality criterion, and even if the recognition of the observed objects, i.e. their partition into different classes, is based on their shape alone. Indeed, the aim is different: in the former case the explorer must collect enough data to *interpolate* the shape/position/pose of an

[2]As usual in least-squares fitting, this uncertainty is quantified by a covariance matrix, obtained by inverting the Hessian of the fit error function at the optimum $\mathbf{m}$. Note that this way of quantifying the uncertainty requires only a *local* Gaussian approximation of the posterior distribution of the models around the optimum. Although the overall posterior hardly ever follows a normal, particularly in the case of nonlinear models, it is generally dominated by a Gaussian peak in a neighborhood of its maximum, on account of both the central limit theorem and the quadratic behaviour of a non-degenerate error function at the optimum.

object as accurately as possible, whereas in the latter the data must be sufficient to *differentiate* the shape of that object from those of the others. Ideally, if there were a viewpoint from which a single disambiguating shape feature could be observed, even with high uncertainty, we should immediately proceed to collect data from there, since this would likely yield a confident recognition.

Consequently, an information based criterion for active data collection, aimed at unambiguously (or, rather, less and less ambiguously) classifying uncertain shapes into different objects, must take into account not only how the new data reduce the uncertainty in the shapes, but also how that reduction in shape uncertainty is effective at reducing the (properly quantified) ambiguity of the classification. This calls for an analysis of how the shape uncertainty is propagated (through the classifier) into the class probability estimates, as well as of the sensitivity of the estimates themselves to changes in the shape parameters.

The particular classification technique used will determine the form of this propagation, although its conclusions should be consistent across different classification methods. For the problems at hand a data-driven approach to classifier design seems mandatory, although simple parametric methods (e.g. Gaussian approximations "in the large") are likely to fail.[3] However, such a probabilistic recognition framework can be cast in terms of more powerful connectionist learning and classification algorithms in a rigorous and structured manner [3, 7, 14], leading to improved classifier performance. The classifier used for the experiments reported in this paper has been implemented with a connectionist architecture, trained and selected using the Bayesian methods for classification networks proposed by MacKay [8] (our experimental results give indications of the adequacy this choice to the task at hand). However, the proposed formalism can be applied to a much wider family of supervised classification methods.

In the following sections we will elaborate on these issues in broader detail. Specifically, in Section 2 we briefly summarize the technique of autonomous exploration for shape reconstruction, upon which our active recognition technique is built. The first main section of the paper, Section 3, describes our model-based classification approach, and how it accounts for model uncertainty. In Section 4 we discuss how to properly measure the ambiguity in a current classification and its expected variation when new data are collected, and introduce the mathematical foundation of an efficient data-gathering methodology for ambiguity reduction. In Section 5 we present experimental demonstrations of these techniques, concluding in Section 6 with a discussion of the results obtained and further research.

2. Autonomous exploration for shape reconstruction

Whaite and Ferrie's [15, 16, 17] technique for actively modelling the shape, pose and position of an object is based on the following idea: when a new datum $\mathbf{x}_{N+1}$ is added to the dataset $\{\mathbf{x}\}$, to which model $\mathcal{M}(\mathbf{x}; \mathbf{m}_N)$ has been fitted, and a new optimal

[3] In the case of superellipsoid models, for example, the inadequacy is due to view-dependent degeneracies [11], that cause the posterior model distribution to variably be uni- or multi-modal, depending on the object and the chosen views.

fit is computed, the shape and the volume of the uncertainty bubble (the ellipsoids of confidence) in the model parameters $\mathbf{m}_N$ shrink respectively as

$$\mathbf{C}_{N+1} \simeq \left(\mathbf{C}_N^{-1} + \beta_\nu \, \mathbf{g}_m \mathbf{g}_m^T\right)^{-1}, \tag{1}$$

$$\det(\mathbf{C}_{N+1}) \simeq \frac{\det(\mathbf{C}_N)}{1 + \beta_\nu \, \mathbf{g}_m^T \, \mathbf{C}_N \, \mathbf{g}_m}, \tag{2}$$

where $\beta_\nu \triangleq 1/\sigma_\nu^2$ is the noise level in the measurements (the noise being assumed additive and i.i.d.), and $\mathbf{g}_m \triangleq \partial\mathcal{M}/\partial\mathbf{m}$ is the sensitivity of the model $\mathcal{M}(\mathbf{x}; \mathbf{m}_N)$ to its parameters, computed in $\mathbf{x}_{N+1}$. These expressions are exact for linear models, and neglect second order terms for nonlinear models. Equation 2 implies that we may reduce the uncertainty either – not surprisingly – by using less noisy sensors, or by taking advantage of the data-dependent term

$$\sigma_m^2 \triangleq \mathbf{g}_m^T \, \mathbf{C}_{N-1} \, \mathbf{g}_m, \tag{3}$$

which is the predicted variance of the interpolant $\mathcal{M}$ at the measurement location. Hence a prescription for efficiently increasing the model accuracy by collecting new data follows:

> *Take further measurements where the predicted errorbars are largest, until they equal the noise level.*

This is the same as Fedorov's "Q-optimal" experiment design [4]. For linear models the sensitivity $\mathbf{g}_m$ depends only on the locations, hence an entire exploration plan can be devised even before taking a single measurement. In the nonlinear case a more conservative attitude is necessary, and feedback from the current model estimate is used to iteratively improve the estimate itself. The above expression will be approximately valid only in a neighborhood of the current $\mathbf{m}_N$, and will depend on its value, hence the prescription weakens as: *Take new measurements where the currently predicted errorbars are largest, in a small neighborhood of the current measurement location, until* etc.[4] Note that this strategy can be described as a descent along the gradient of the current uncertainty in the model, but it is *very* different from "gradient descent" in the usual sense of this expression, i.e. descent along the gradient of a fixed cost function. Here the cost function changes at every step because of the new measurements, with the net effect of "pushing" the explorer away from already visited viewpoints. This effectively overcomes a major drawback of "gradient descent", namely faltering on local minima.

In the present implementation of the autonomous explorer, the measurements are 3D coordinates of object surface points, measured with a laser rangefinder probe mounted on a variety of robot devices, and the – highly nonlinear – models used are superellipsoids [1] fitted via least-squares optimization to the segmented parts of the object.[5] The above prescription specifies where *on the object* the next measurement

[4] Strictly speaking, we should say "in a small neighborhood of the current model estimate $\mathbf{m}_N$": a continuity argument is necessarily involved here.

[5] The parameters $\mathbf{m}$ are in this case the parts' breadth, height and width, Euler angles of their principal frame, displacement coordinates, the two superellipsoid ϵ parameters defining the "squareness"/"roundness" of the shape.

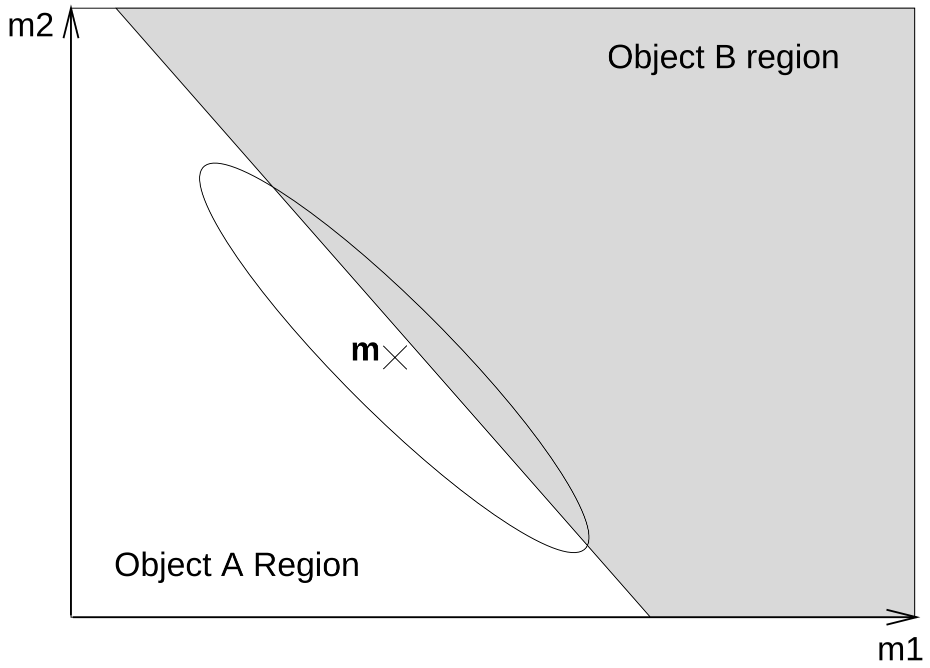


FIGURE 3. An example showing why reducing model uncertainty does not necessarily lead quickly to confident recognition.

should be made. The corresponding sensor location is found by solving an inverse ray-tracing problem, which is not trivial due to possible degeneracies in the points-to-sensor mapping. Details of the solution can be found in the above references.
We can, at this point, better clarify our previous argument about why this sensor planning strategy is *not* optimal when the task is to recognize confidently the objects through their models, and not to simply model the objects as accurately as possible. Consider, for instance, a recognition problem in the simple model space depicted in Figure 3. The models here have two components, hence the model's confidence ellipsoids are really ellipses, the covariance matrix $\mathbf{C}$ being 2×2. A current model estimate $\mathbf{m}$ has been computed, and one of eigenvalues of $\mathbf{C}$ largely dominates the other: this is a common situation, since early measurements usually constrain some parameters much better than others (e.g. "height" and "width" more than "depth", for volumetric models). Now, the larger eigenvalue will give a dominant contribution to the model's predicted variance σ_m^2 over all $\{\mathbf{x}\}$ space, since the $\mathbf{x}$ locations where the sensitivity $\mathbf{g}_m$ is orthogonal to the larger eigenvector yield small variances anyway. Hence the above planning strategy will, at least initially, drive the explorer where the measurements best reduce the larger eigenvalue. But this is certainly inefficient for recognition purposes if, as in the figure, a line separating the object classes in model space (either deterministicaly or as a Bayesian decision limit) lies almost parallel to the dominating eigenvector. What we really need to know to recognize the object is how far the model is from that line, hence what matters in this case is precisely the contribution to model uncertainty due to the *smaller* eigenvalue. Note that the relative sizes of the confidence eigenvalues are determined by the relation between

measurements and models, while the (probabilistic) recognition is determined by the relation between (uncertain) models and classes, and the two relations are clearly not the same, although we do expect that an accurate model estimate should produce a confident recognition. Evidently we are not interested in model accuracy *per se*, but rather in correlating model accuracy with the partition of models into object classes. This "correlation" idea, and how to exploit it to advantage, is the main point of Section 4, and the principal contribution of this paper.

3. Object classification

The first problem to solve is then how to properly use uncertain models for the purposes of recognizing the modelled objects. Following the discussion in the introductory section, from now on we shall systematically (ab)use the term recognition to mean a probabilistic classification of a given model object, i.e. a conditional probability distribution $P(O_i|\mathbf{m})$ over q possible objects $\{O_i\}_{i=1,..,q}$ given the current uncertain object shape, with the Bayesian interpretation of probability as a measure of the state of knowledge (or ignorance). The "forward" recognition problem of the scheme in Figure 2 is then reduced to designing a probabilistic classifier of the uncertain shape data with respect to a known set of objects.

In our current implementation, this problem is solved by a connectionist architecture. For sake of generality, however, we shall simply require that the classifier be of the discriminative type, i.e. represented by some function $\mathbf{f}(\cdot)$ of the model parameters, providing an estimate of the probability $P(O_i|\mathbf{m})$ that a model parameter vector $\mathbf{m}$ belongs to object class O_i:

$$P(O_i|\mathbf{m}) \simeq f_i(\mathbf{m};\mathbf{w}) \; , \tag{4}$$

where $f_i(\cdot)$ is the i-th component of $\mathbf{f}(\cdot)$, and $\mathbf{w}$ is the vector of the internal parameters of the classifier. It is convenient to express $\mathbf{f}(\cdot)$ in the form,

$$f_i(\mathbf{m};\mathbf{w}) = f_i(\mathbf{a}(\mathbf{m};\mathbf{w})), \tag{5}$$

which embodies the internal structure of the classifier in the function $\mathbf{a}(\cdot)$, whose values are translated into probabilities by $\mathbf{f}(\cdot)$. Two forms are commonly used for $\mathbf{f}(\cdot)$: the "logistic" function $f(a) \triangleq e^a/(1+e^a)$ for the binary classification case,[6] and the "generalized logistic" or "softmax" function $f_i(\mathbf{a}) \triangleq e^{a_i}/\sum_j e^{a_j}$ for the multiple class case. The appropriateness of these functions for the output of a classifier is well known to statisticians: they are *the* functions that convert probability ratios into probabilities.

The classifier parameters $\mathbf{w}$ of a particular classifier structure are optimized by means of a training set of labelled data. The training procedure and the structure selection technique we use closely follow the hierarchical Bayesian methods proposed by MacKay in his seminal works of 1991–92 [6, 7, 8]. Among the leading ideas of MacKay's approach is that a supervised classifier architecture should be regularized

[6]Both $\mathbf{f}(\cdot)$ and $\mathbf{a}(\cdot)$ become scalar functions in the binary case: a classifier with only one output is used, whose value is interpreted as an estimate of the conditional probability of one class, or one minus the probability for the other.

and selected by optimizing its "evidence" as supported by the available data, i.e. its likelihood marginalized with respect to all possible parameter vectors $\mathbf{w}$.
Our approach, however, differs from MacKay's in that we must use the classifier to make predictions in presence of uncertainty in the inputs (in our case, the uncertain models $\mathbf{m}$). As far as learning is concerned, this alone would make the problem hardly tractable: Bayesian interpolation of data with noise in the inputs is notoriously difficult, and solutions have so far been proposed for the linear case only [5]. Yet we can easily (although not cheaply) negotiate this difficulty for training purposes, since we can use the shape-driven explorer, and merge data from multiple views until the model uncertainty is made negligible. We can afford the computational cost of this procedure of data collection because the classifier is trained off line, once and for all for a given set of possible objects.
However, in order to efficiently *use* an already trained classifier to make predictions in a recognition-driven exploration, we want to merge as few views as possible. We must then account for the uncertainty because, in the case of significant model error bars $\mathbf{C}$, the classifier output to the most probable model $\mathbf{m}$ is generally biased, i.e. *not* a good representative of the class for the posterior distribution of possible models $\boldsymbol{\mu}$ given the data, $P(\boldsymbol{\mu}|\{\mathbf{x}\})$, even for $\boldsymbol{\mu}$ in the neighborhood of $\mathbf{m}$.[7] The correct approach is to treat the models as nuisance parameters to be "marginalized away", so to expose the data's support for each class:

$$P(O_i|\{\mathbf{x}\}) = \int d\boldsymbol{\mu}\, P(O_i|\boldsymbol{\mu})\, P(\boldsymbol{\mu}|\{\mathbf{x}\}) \quad , \tag{6}$$

where $P(O_i|\boldsymbol{\mu})$ is given by the trained classifier's output, and the integral is extended over model space. This integral can be computed using the following approximations:

- The model posterior $P(\boldsymbol{\mu}|\{\mathbf{x}\})$ is locally approximated, in the neighborhood of the optimal fit $\mathbf{m}$, with a Gaussian with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$.
- The classifier's function $\mathbf{a}(\cdot)$ is locally linearized in the neighborhood of $\mathbf{m}$, i.e. $\mathbf{a}(\boldsymbol{\mu}) \simeq \mathbf{a}(\mathbf{m}) + \mathbf{g_a}(\boldsymbol{\mu} - \mathbf{m})$, where $\mathbf{g_a} \triangleq (\partial \mathbf{a}/\partial \boldsymbol{\mu})_{\boldsymbol{\mu}=\mathbf{m}}$ is the sensitivity of $\mathbf{a}(\cdot)$ computed at $\mathbf{m}$.

These imply that the distribution of the $\mathbf{a}(\cdot)$ values can be locally approximated, in a the neighborhood of $\mathbf{a}(\mathbf{m})$, by a Gaussian $N(\mathbf{a}; \mathbf{a}_m, \mathbf{C}_a)$, with mean $\mathbf{a}_m \triangleq \mathbf{a}(\mathbf{m})$ and covariance matrix $\mathbf{C_a} \triangleq \mathbf{g}_a^T\, \mathbf{C}\, \mathbf{g}_a$, yielding for the marginalized prediction,

$$P(O_i|\mathbf{m}, \mathbf{C}) \triangleq P(O_i|\{\mathbf{x}\}) \simeq \int d\mathbf{a}\, f_i(\mathbf{a})\, N(\mathbf{a}; \mathbf{a}_m, \mathbf{C}_a)\,, \tag{7}$$

the integral being now extended over $\mathbf{a}$ space. The symbol $P(O_i|\mathbf{m}, \mathbf{C})$ for the marginal, although an abuse of notation ($\mathbf{m}$ and $\mathbf{C}$ are not random variables), is used to remark that the approximation can be computed using model-related quantities only. This approximation is adequate as long as $\mathbf{a}(\cdot)$ does not vary too rapidly within the model's confidence interval. The integral of a logistic (or softmax) function times a Gaussian cannot be solved in closed form, though it can easily be evaluated numerically. However, a well behaved approximation has been given[12] for the logistic case

[7] Technically, this is a consequence of the normalization of probabilities, which causes the mapping from models to class probabilities to be necessarily nonlinear.

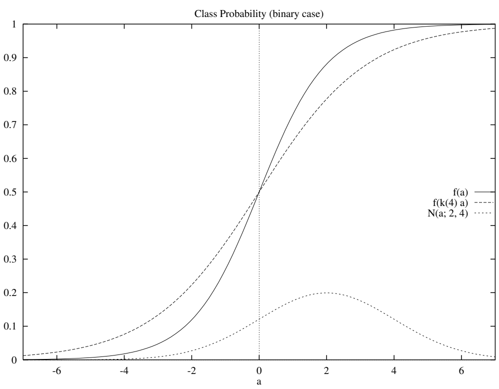
(i.e. binary classification), from which useful insight can be drawn,

$$P(O_i|\mathbf{m}, \mathbf{C}) \simeq \phi(a_m, s^2) \triangleq f\left(a_m \; k(s^2)\right) \quad , \tag{8}$$

where $s^2 = \mathbf{g}_a^T \, \mathbf{C} \, \mathbf{g}_a$ is the (scalar, in this case) variance of $a(\cdot)$, and

$$k(s^2) \triangleq \left(1 + \pi s^2/8\right)^{-\frac{1}{2}} \quad . \tag{9}$$

This result is intuitively satisfactory: $k(s^2) \in [0, 1]$, hence the uncertainty in the input reduces the confidence in the prediction by "pulling" the argument of $f(\cdot)$ toward 0, and hence the estimated class probability toward $\frac{1}{2}$. This is demonstrated in Figure 4 (left). Note also that marginalization cannot change the "direction" of the classification: if the non-marginalized probability is greater than $\frac{1}{2}$, so is the marginalized, hence the operation is pointless for the gambler's "all-or-none" decision policy, which assigns a model to the class having largest probability (when decision risks are symmetric). However, it *does* make a difference if a more sophisticated policy can be used, for example one in which a decision is delayed, and more data are collected, whenever the classification confidence is below a certain threshold. This is exactly the class of policies to which the present work may be of interest.
An example of the usefulness of this technique is also shown in Figure 4 (right). The two curves represent the marginalized and non-marginalized classifier prediction, i.e. the probabilities of being object R given the model alone, $P(R|\mathbf{m})$, and given the model and its uncertainty, $P(R|\mathbf{m}, \mathbf{C})$, at each step of a recognition-driven exploration around a test object which actually was an R (for Round; details of the exact experimental setup will be given shortly). Note how the non-marginalized prediction is initially very confident, but wrong, while the marginalized prediction becomes confident, and correct, only when enough data have been collected, and the uncertainty bubble in the shape $\mathbf{m}$ has shrunk enough to support such confidence.



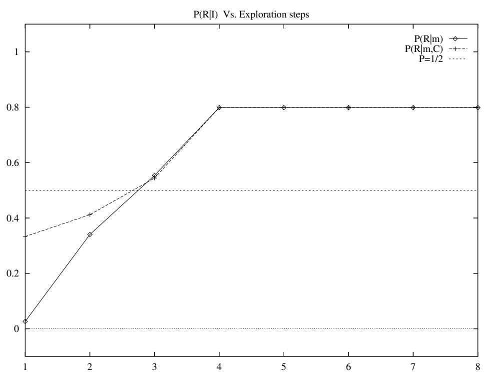


FIGURE 4. Left: examples of marginalized and non-marginalized predictions in the binary case for $s^2 = 4$. Right: comparison between marginalized and non-marginalized class predictions, respectively $P(R|\mathbf{m}, \mathbf{C})$ and $P(R|\mathbf{m})$, in the course of a recognition-driven exploration around an object of R(ound) type.

The above approximation falls graciously for $a^2 \gg s^2$, and is formally identical to MacKay's [8]. The difference is in the uncertain variables the marginalization is made over: the classifier parameters $\mathbf{w}$ in his case, the classifier inputs $\mathbf{m}$ in ours.
These marginalized, or "moderated" predictions are used in the exploration technique proposed in the next section.

4. Autonomous active recognition

Now to the main point of this paper: finding a criterion to efficiently collect new data in order to reduce the recognition ambiguity, thus defining a recognition-driven planning strategy for an autonomous explorer.
We first need to select a measure of the reduction in classification ambiguity when the uncertainty about the model is changed by gathering new data. A natural candidate is the variation in the Shannon entropy of the marginal class distribution defined in the previous section,

$$H(P(O_i|\mathbf{m},\mathbf{C})) = \sum_i P(O_i|\mathbf{m},\mathbf{C}) \log \frac{1}{P(O_i|\mathbf{m},\mathbf{C})} \quad , \tag{10}$$

and we shall define the information gain across successive measurements as $\Delta H \triangleq H_N - H_{N+1}$, so that a positive gain corresponds to a reduction of classification uncertainty. Now we can generally expect to gain information by taking measurements: elementary information theory shows that conditioning on average reduces the entropy of a distribution (data are useful, after all). But we are interested in finding those locations where measurement yields *large*, ideally maximal, information gains.
Let us consider the binary case again. Using the above approximations, the sensitivity of the entropy to the variance of a in the neighborhood of $a_m = a(\mathbf{m})$ is

$$\begin{aligned} \frac{\partial H}{\partial(s^2)} &= -\frac{\partial}{\partial(s^2)}[\phi \log \phi + (1-\phi)\log \phi] \\ &= [\log(1-\phi) - \log \phi]\, \phi\, (1-\phi) \\ &\quad \times a\ dk/d(s^2)\ , \end{aligned} \tag{11}$$

where $\phi = \phi(a, s^2)$, and use is made of the identity $f'(x) = f(x)\,(1 - f(x))$, valid for the logistic function. But

$$dk/d(s^2) = -(\pi/8)\,(1 + s\,\pi/8)^{-3/2} < 0 \quad , \tag{12}$$

and the term $[\log(1-\phi) - \log\phi]$ in Equation 11 obviously has sign opposite to that of a. Hence, within the approximation, $\partial H/\partial(s^2)$ is always positive: lowering s^2 implies lowering H, so we always gain information when the uncertainty in a is reduced. We then focus on the properties of the inputs $\mathbf{m}$ that affect a's uncertainty. From the identity

$$\left(\mathbf{C}_N^{-1} + \beta_\nu \mathbf{g}_m \mathbf{g}_m^T\right)^{-1} = \mathbf{C}_N - \frac{\beta_\nu \mathbf{C}_N \mathbf{g}_m \mathbf{g}_m^T \mathbf{C}_N}{1 + \beta_\nu \mathbf{g}_m^T \mathbf{C}_N \mathbf{g}_m} \quad , \tag{13}$$

and from Equation 1 we have, after collecting a new datum $\mathbf{x}_{N+1}$ and refitting the model, using again a local linear approximation for the classifier (i.e. almost constant $\mathbf{g}_a \simeq \mathbf{g}_a(\mathbf{m})$) ,

$$
\begin{aligned}
s_{N+1}^2 &= \mathbf{g}_a^T\,\mathbf{C}_{N+1}\,\mathbf{g}_a = \mathbf{g}_a^T\,\mathbf{C}_N\,\mathbf{g}_a - \mathbf{g}_a^T\left(\frac{\beta_\nu\,\mathbf{C}_N\,\mathbf{g}_m\,\mathbf{g}_m^T\,\mathbf{C}_N}{1+\beta_\nu\mathbf{g}_m^T\,\mathbf{C}_N\,\mathbf{g}_m}\right)\mathbf{g}_a \\
&= s_N^2 - \frac{\beta_\nu\,c_{am}^2}{1+\beta_\nu\,\sigma_m^2} \quad .
\end{aligned} \tag{14}
$$

In this equation the scalar term,

$$c_{am} \triangleq \mathbf{g}_a^T\,\mathbf{C}_N\,\mathbf{g}_m \tag{15}$$

measures the correlation between the sensitivities of a and $\mathcal{M}$ to the model parameters, in the metric defined by the current uncertainty about the parameters themselves. Here we have an intuitively appealing result: the reduction in recognition ambiguity is driven not only by a decrease in the model uncertainty σ_m^2, but also by an increase in the squared correlation c_{am}^2 between the incremental change in the model when new data are gathered, $\mathbf{g}_m$, and the corresponding incremental change in the classifier's idea of the model itself, $\mathbf{g}_a$.

The fact that this measure is squared in the above expression means that negative correlations count as much as positive ones in reducing the ambiguity. And understandably so: taking new data where the correlation is negative implies a (predicted) decrease of the current confidence level in the belonging of the modelled object to a certain class, hence an increase of the confidence that it belongs to another one. This may be exactly the right thing to do if, after taking new data, the object starts looking as those of the latter class. And if it does not, information has been gained anyway by ruling out that possibility. Hence the squared sensitivity correlation acts in the sense of "marking" measurement locations *on the model* which are useful to disambiguate the objects, while the model's variance σ_m^2 downweights this tendency by the uncertainty about the where the locations themselves are, given the current uncertain model. As in sec. 2, we stress that finding the sensor positions from which the above model locations can actually be measured is a different, implementation-related problem, to be solved, say, using inverse optics for camera/laser based sensors, inverse kinematics for a touch probe, etc.

Note now that, in Equation 14, it is $(\beta_\nu\,\sigma_m^2) \geq 1$, since the model usually cannot be predicted better than it can be measured. If we are in a moderately low-noise (high β_ν) regime, we can assume $(\beta_\nu\,\sigma_m^2) \gg 1$, and then approximate the fraction in that equation with the ratio c_{am}^2/σ_m^2. The advantage is that we can compute this quantity without having to actually estimate the noise level itself, since it depends only on the current model and on the classifier.

A new prescription for recognition-driven autonomous exploration then follows:

> *Take further measurements at locations that maximize the ratio of the squared correlation between class and model sensitivities (in the metric defined by the current model's uncertainty) to the predicted variance of the model, until the classification entropy stops decreasing.*

The same argument of section 2 for implementing this strategy in an incremental fashion for the nonlinear model case obviously applies here, as does the note in the same section about the unconventional "gradient descent" that such strategies implement.

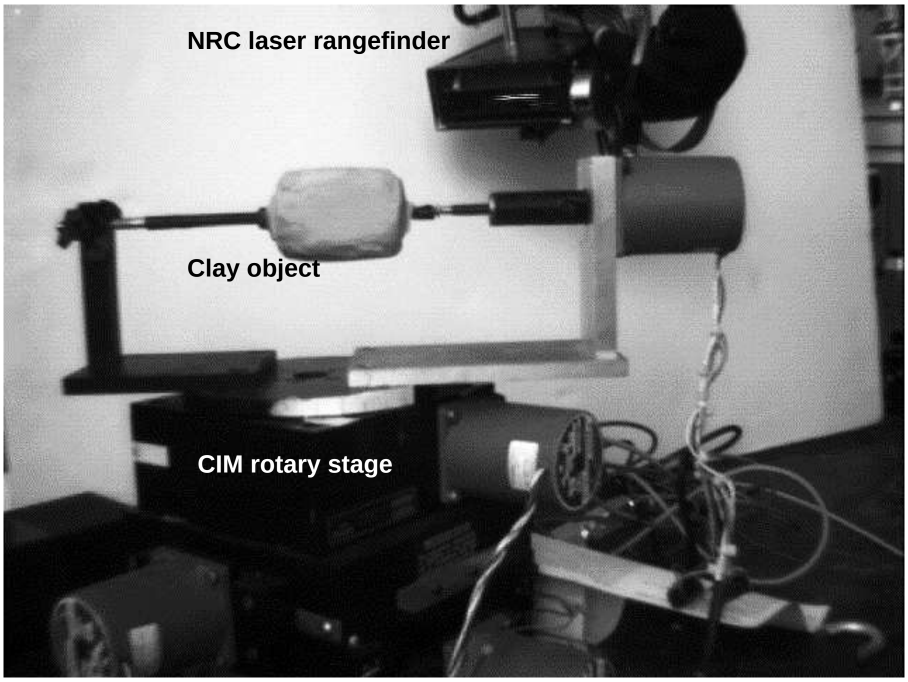


FIGURE 5. The experimental setup used to test the proposed recognition-driven exploration strategy.

5. DEMONSTRATIONS

The recognition-oriented exploration strategy introduced in the previous section is demonstrated with a series of sensor-planning experiments. The experimental setup, portrayed in Figure 5, consists of a rotary stage with four degrees of freedom holding objects to be analyzed, and of a two-axis laser rangefinder probe mounted on the end effector of a PUMA 560 manipulator. The controlling software is based on the previous Whaite and Ferrie's autonomous modeller/explorer [15, 16, 17], augmented with a purposely designed Bayesian connectionist simulator and auxiliary software in order to implement the new exploration strategy (class probability estimation, class sensitivity computation, exploration step planning, performance statistics collection, etc). The robot control system effectively moves the probe along the geodesic lines of a "viewsphere" centered on the object being analyzed, while the higher level planning software decides the direction and size of each geodesic step. The size of the processed depth images was 40×40 pixels, covering an area on the target of $15 \times 15\,cm^2$ circa, with uniform resolution.

A binary recognition problem has been defined by considering a set of 20 single-part objects, modelled with clay according to two basic shapes: "round" (R) objects,

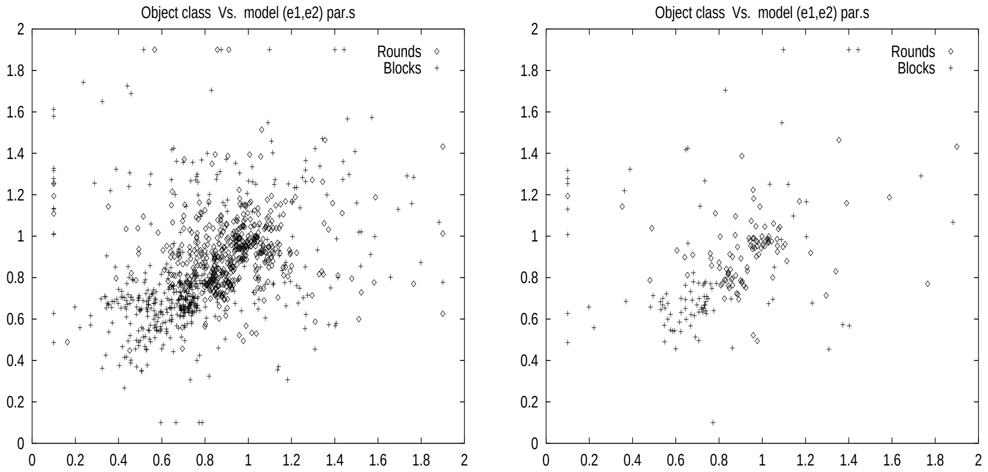


FIGURE 6. Both graphs are slices on the (ϵ_1, ϵ_2) plane of the 11-dimensional superellipsoid parameter space. Left: whole model set collected via shape-oriented exploration (1000 points). Right: the models recovered at the last exploration steps, used to train the classifier (200 points).

suitable to be approximated by superellipsoids with ϵ parameters in the neighborhood of 1, and "block" (B) objects, for which ϵ parameters close to 0 can be expected. However, the basic shapes were heavily deformed, in order to make the classification problem non trivial. In order to test the robustness of the classifier, the labelling of the objects as B's or R's is done on the basis of the experimenter's personal judgement, prior to seeing the actual values of the model parameters. The data used to train the classifier have been collected by running 10 model-oriented exploration per object, each initialized on a random point of the viewsphere and run for 5 steps. For these object shapes it can be expected that the model separation in the R and B classes be maximally evident in the (ϵ_1, ϵ_2) plane of the 11-dimensional parameter space. The actual model distribution per class on that plane for all the models recovered during all explorations is then shown in the left-hand side graph of Figure 6: it is evident that a clear-cut class separation is not feasible (given also the uncertainty in the parameters themselves), and a statistical approach seems necessary. The graph on the right-hand side of the same figure shows the parameters recovered after merging all the views (i.e. at the last exploration step): these ones, following the discussion in Section 3, are the only actually used for training purposes.
A set of feedforward two-layer (one hidden) connectionist classifiers has been trained on this classification problem on the 200 points of training set, using MacKay's "evidence" framework, and an 11-6-1 network characterized by the highest evidence has been selected for the purposes of the exploration. The agreement between evidence and empirical performance for this network is also supported by the percent error in "all-or-none" classifications on the whole (all views) data set, which is 19.8%: not a bad result, considering that the majority of the set consists of shape models built early in the exploration, hence affected by high uncertainty (and usually wrong). The

final error on the training set is 4.2%. The classification function computed by this network is shown in Figure 7, again in the (ϵ_1, ϵ_2) plane only.
The performance of the proposed recognition-oriented exploration strategy has been comparatively assessed on a new object set, composed of five B and five R objects. Three ten-step explorations have been run for each of them, respectively using the new strategy, Whaite and Ferrie's model-oriented strategy, and a random walk strategy.The comparison with the second strategy is made to verify whether our strategy, which does not try to accurately model the shape, is actually effective even if the recognition can be based only on the shape of the objects (or, rather, on the differences in their shapes). The last strategy is used for a worst-case control test, and consists of simply stepping the probe in a random direction along a geodesic from the current position. The total number of exploration tests performed on all objects and with all strategies is 90.

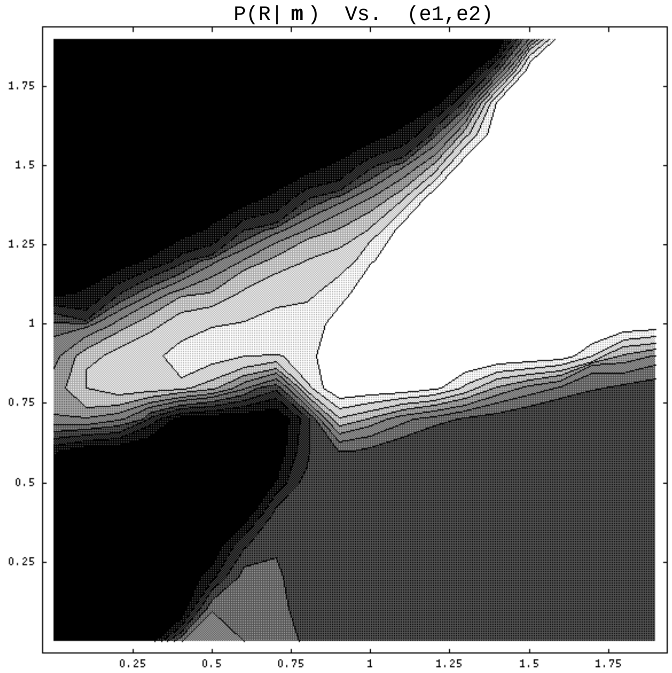


FIGURE 7. Graph (slice on the (ϵ_1, ϵ_2) plane) of the non-moderated classification function $P(R|\mathbf{m})$ learned by the selected network after training (white indicates probability one of being and R object).

The test data thus collected are evaluated according to the following criteria:

- An exploration is scored as having produced a confident and stable recognition (CSR) if class probability estimates in excess of 0.75 are maintained for at least three consecutive exploration steps. Unstable recognitions (usually due to non stable models for heavily deformed shapes), and stable recognition with non confident probability values are noted separately, and counted as rejections.
- For the CSRs, the recognition accuracy is evaluated. This is simply defined as 1 minus the absolute difference between the explorer's class probability estimate

and its true value (1 or 0, according to the *prior* judgement of the experimenter who prepared the object). The accuracy is therefore close to 1 if the explorer assigned a large probability to the correct class, and close to 0 otherwise.
- For the CSRs, the recognition time is also evaluated. This is defined as the number of steps performed before the explorer's class probability estimate becomes bounded above 0.75.

The observed CSR/rejection, accuracy and recognition time figures are reported in Figure 8, respectively in Tables 1, 2 and 3.
These can be interpreted as follows:

- The number of rejections is small for all strategies. This is indicative of good stability in the model builder and of good robustness in the classifier, when the system is confronted with random object deformations and noise.
- The accuracy is the same for all strategies, within the experimental error. This confirms the previous conclusion from a different point of view: the recognition is stable and accurate, as long as enough data are collected, either in a "smart" way by following a strategy, or by simply wandering at random on the viewsphere.
- The interesting result is the recognition time. Given the simplicity of the objects used so far, the relevant point here is actually not the mean number of steps, although this indicates that the new strategy is on average one step ahead, but the small spread of this value compared with the other ones. It indicates that the newly proposed recognition-driven strategy is consistently faster: it can be reasonably expected to provide a good class estimate within 2 steps.

1	Recognition	Model	Random
CSR	88	84	86
REJ	2	6	4

2	Recognition	Model	Random
Mean accuracy	0.92	0.84	0.88
Std. dev.	0.23	0.33	0.28

3	Recognition	Model	Random
Mean time	2.03	2.96	2.67
Std. dev.	0.46	2.05	1.87

FIGURE 8. Summary of the results for the three tested exploration strategies. Table 1 (top): Number of confident stable recognitions (CSR) and rejections (REJ) on all the performed test explorations. Table 2 (middle): Mean accuracies and their standard deviations. Table 3 (bottom): Mean exploration times (in number of steps) and their standard deviations.

This last conclusion agrees with the experimental observation that the new strategy seems to act as to quickly drive the probe to viewsphere locations from where "corners" can be observed when the object is a block (B). In other words, it seems to indicate that the explorer uses correctly the "contextual" information coming from the classifier, i.e. that the shapes produced from the modeller can be either blocks or rounds, and nothing else, thus allowing to quickly confirm or refute the current guess about the class by checking whether the distinguishing shape features, corners, are actually there or not. Where is "there", i.e. where the corners, if any, are likely to be is jointly determined by the modeller producing $\mathbf{m}$ and the corresponding predictions σ_m^2 about the object surface, and by the classifier indicating, via the c_{am}^2 values, how much those predictions agree or disagree with the current class estimate.

6. Discussion

The results presented, although preliminary, show that our recognition-oriented sensor planning strategy can effectively take advantage of the *difference* in shape among the objects. The local approximations used in the theoretical analysis (local Gaussian approximation of the model posterior, local linearization of the classifier, approximate marginalization of the class estimates) seem adequate in the globally nonlinear case experimented (superellipsoid models). More extensive tests, using more complex shapes, are currently being performed.

The theory underlying the proposed strategy is strictly valid for binary recognition problems. Yet the binary hypothesis has been made only in order to reach a closed-form solution for the marginalized predictions, through suitable approximations. However, our main result is derived in Equations 14, and the first two of them do not depend on that hypothesis. They thus show that the idea of looking for locations on the model which are highly correlated with one class (or a subset of the possible classes), and have small predicted variance, is a more fundamental one. Note that in the multi-class case the term $\mathbf{g}_a$ is to be reinterpreted as the Jacobian matrix of $\mathbf{a}(\cdot)$, and consequently s^2 as a covariance matrix. Work is under way to formulate an appropriate generalization.

Acknowledgments

The authors wish to thank Gilbert Soucy and Peter Whaite for providing part of the software used in the simulations.

References

[1] A. H. Barr. Superquadrics and angle preserving transformations. *IEEE Computer Graphics and Applications*, 1(1):11–23, Jan. 1981.

[2] R. Bolles. On optimally combining pieces of information, with application to estimating 3-d complex-object position from range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(5):619–638, 1986.

[3] J. S. Bridle. Probabilistic interpretation of feedforward classification networks outputs, with relationship to statistical pattern recognition. In F. Fogelman-Soulié and J. Hérault, editors, *NATO ASI*, volume F68. Springer-Verlag, Berlin, 1990.

[4] V. Fedorov. *Theory of optimal experiments.* Academic Press, New York, 1972.

[5] S. F. Gull. Bayesian data analysis:straight-line fitting. In J. Skilling, editor, *Maximum entropy and Bayesian methods*, pages 511–518. Kluwer Academic Publishers, 1989.

[6] D. J. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):589–603, 1991.

[7] D. J. MacKay. A practical Bayesian framework for backprop networks. *Neural Computation*, 4(3):448–472, 1991.

[8] D. J. MacKay. *Bayesian methods for adaptive models.* PhD thesis, California Institute of Technology, Pasadena, CA, USA, 1992.

[9] P. MacKenzie and G. Dudek. Precise positioning using model-based maps. In *Proceedings of the International Conference on Robotics and Automation*, San Diego, CA, 1994. IEEE Press.

[10] A. Pentland and S. Sclaroff. Closed form solutions for physically based shape modelling and recognition. In T. Kanade and K. Ikeuchi, editors, *IEEE Transactions on Pattern Analysis and Machine Intelligence: Special Issue on Physical Modeling in Computer Vision*, volume 13(7), pages 715–729, July 1991.

[11] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(2):131–147, Feb. 1990.

[12] D. J. Spiegelharter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *NETWORKS*, 20:579–605, 1990.

[13] J. Subrahmonia, D. B. Cooper, and D. Keren. Practical reliable bayesian recognition of 2D and 3D objects using implicit polynomials and algebraic invariants. LEMS 107, Brown University LEMS, Laboratory fo Engineering Man/Machine systems, Division of Engineering, Brown University, Providence, RI 02912, USA, 1992.

[14] N. Tishby, E. Levin, and S. Solla. Consistent inference of probabilities in layered networks. In *Proc. Internat. Joint Conf. on Neural Networks*, Washington, 1989.

[15] P. Whaite and F. P. Ferrie. From uncertainty to visual exploration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1038–1049, Oct. 1991.

[16] P. Whaite and F. P. Ferrie. Active exploration: Knowing when we're wrong. In *PROC. Fourth International Conference on Computer Vision*, pages 41–48, Berlin, Germany, May 11-14 1993. Computer Society of the IEEE, IEEE Computer Society Press.

[17] P. Whaite and F. P. Ferrie. Autonomous exploration: Driven by uncertainty. In *Proceedings, Conference on Computer Vision and Pattern Recognition*, pages 339–346, Seattle, Washington, June 21-23 1994. Computer Society of the IEEE, IEEE Computer Society Press.

Centre for Intelligent Machines, McGill University, 3480 University St., Montréal, Que., Canada, H3A 2A7
E-mail address: `franco@cim.mcgill.ca, ferrie@cim.mcgill.ca`