Active sampling strategies for multihypothesis testing

Stéphane HERBIN

ONERA
Département Traitement de l'Information et Modélisation
29, avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex
France
Stephane.Herbin@onera.fr
phone: (33)+1 01 46 73 49 91
fax: (33)+1 01 46 73 41 67

This paper presents a rationale for the design of optimal sequential sampling procedures for multi-hypothesis discrimination where a system selectively queries the environment based on the current state of the discrimination process.
The environment is modelled as a controlled i.i.d. process conditionned by various hypotheses. Recognition is achieved when the test identifies the correct hypothesis describing the environment behavior.
As the testing proceeds, hypotheses may be rejected with infinite confidence when feature values with zero probability are observed. The sampling strategy is stationary but is updated each time a hypothesis is rejected. It is chosen according to a criterion measuring the recognition error speed of convergence to zero when the number of samples goes to infinity. This criterion is obtained by Large Deviation Theory techniques and characterizes globally the multihypothesis discrimination problem.
An application on 2D rotation invariant shape recognition with non-closed noisy contours illustrates the approach.

Keywords : multihypothesis decision process, optimal sampling strategies, large deviations theory, shape recognition.

This paper is presented to EMMCVPR 2003 only.

Active sampling strategies for multihypothesis testing

Stéphane HERBIN

ONERA
Département Traitement de l'Information et Modélisation
29, avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex
France
Stephane.Herbin@onera.fr

Abstract. This paper presents a rationale for the design of optimal sequential sampling procedures for multi-hypothesis discrimination where a system selectively queries the environment based on the current state of the discrimination process.
The environment is modelled as a controlled i.i.d. process conditionned by various hypotheses. Recognition is achieved when the test identifies the correct hypothesis describing the environment behavior.
As the testing proceeds, hypotheses may be rejected with infinite confidence when feature values with zero probability are observed. The sampling strategy is stationary but is updated each time a hypothesis is rejected. It is chosen according to a criterion measuring the recognition error speed of convergence to zero when the number of samples goes to infinity. This criterion is obtained by Large Deviation Theory techniques and characterizes globally the multi-hypothesis discrimination problem.
An application on 2D rotation invariant shape recognition with non-closed noisy contours illustrates the approach.
Keywords : multihypothesis decision process, optimal sampling strategies, large deviations theory, shape recognition.

1 Introduction

1.1 The "active" approach

The principle of an "active" testing or recognition approach is to identify the behavior of a possibly dynamic and random environment and assign it to a predefined set of models or labels.

There are several reasons for using an active approach for recognition: a first one is philosophical, since we believe that the true origin of the intelligence of natural systems such as animals or humans lies in the interactive control of their environment, which must be described as a genuine time-dependent process. This question will not be addressed further in this paper.

A second reason, more engineering oriented, is the limitations of the ability of the algorithms embedded in artificial systems to deal with complex objects or environments. Most of the time, the usual approach consists in capitalizing all possible data about the environment and, in a posterior phase, try to make it informative by reducing or transforming it in a goal oriented way.

The informative step is often produced either through the participation of human expertise or knowledge, or automated. In this last option, the design is either constrained by the availability of a huge number of pre-identified data or limited to a small number of hypotheses. However, interesting problems must handle a significant number of hypotheses represented by a small number of exemplars.

We believe that one of the key problems to overcome these limitations is the dynamic management of informative data. Indeed, there is no such a thing as a universal set of object characteristics able to discriminate a priori all kinds of hypotheses. The idea of an active testing approach is to control *on line* in a goal oriented way the choice of the useful features based on a current state of achievement.

1.2 Objectives

We investigate in this paper the optimal design of decision processes based on sequentially querying the values of several features or measures. The testing procedure is said to be "active" since the querying depends on the past collected data.

At each time T, a query or action a_T is generated towards the environment which returns a feature value s_T. All those variables may be random and time dependent.

Given a sequence of action/measurements $\Phi_T = (a_1, s_1, a_2, s_2, \dots a_T, s_T)$ specific to a hypothesis, the procedure of active testing is based on the exploitation of the following recursion:

$$\mathrm{P}[\Phi_T] = \mathrm{P}[s_T \,|\, a_T, \Phi_{T-1}]\mathrm{P}[a_T \,|\, \Phi_{T-1}]\mathrm{P}[\Phi_{T-1}] \tag{1}$$

meaning that the likelihood of observing a sequence Φ_T depends on the past likelihood and on the new measurement.

Once the feature values have been collected, a final decision $B(\Phi_T)$ will identify or reject a hypothesis among a predefined set $\Omega = \{\omega_k\}_{k=1}^N$, or decide to generate new actions to improve the decision safety. We will not deal in the following with this last choice, and take T to be a non random stopping time.

Designing an active decision process consists in solving three problems:

1. Model the action-conditioned probability of observing a feature value given the past: $\mathrm{P}[s_T \,|\, a_T, \Phi_{T-1}]$
2. Define the sequence of actions or queries to be generated : $\mathrm{P}[a_T \,|\, \Phi_{T-1}]$
3. Define a decision procedure based on the collected data : $B(\Phi_T) \in \Omega$

The goal of this paper is to propose several directions to solve those three questions under given assumptions.

1.3 Paper organization

Section 2 discusses the choice of a controlled i.i.d. process for modelling the interactions between the recognition system and the environment. Section 3 describes the structure of the selective testing procedure based on rejecting hypothesis with null likelihood. The description of the on-line sampling strategy based on Large Deviations exponential error rates is presented in 4. An application on noisy shape recognition exploiting the distribution of pairs of contour points is described in 5. Several mathematical results and sketched proofs about large deviations techniques are summarized in the appendices A.1 and A.2.

1.4 Related work

Many recognition algorithms can be described in the "active testing" framework. The differences depend on the choices that have been made to model the environment.

Classification trees [1, 2], which are inherently sequential algorithms, may be the of closest type. They model functional dependence to the past collected feature values as a branch and assign a query to each internal nodes. In their usual setting, however, actions generated and return feature values are deterministic. Similarly to the approach taken in this paper, the design of the querying strategy follows a local optimal construction by minimizing a hypothesis scattering function such as the entropy.

Large deviation theory applied to pattern recognition has been used on a few papers; the main issues have usually been to design bounds able to control the behavior of empirical learning processes [3]. In computer vision, several authors have applied basic large deviation theory results to the analysis of specific algorithms. [4] studies several numerical quantities able to characterize the detectability of simple objects such as curves in a noisy image. [5] uses large deviation theory and statistical mechanics concepts to characterize texture discrimination.

This work is a continuation of previous studies on active recognition [6] and application of Large Deviation Theory techniques to 3D object aspect graph comparison [7] and texture similarity measures [8].

2 Controlled i.i.d processes

This section examines the type of model used to describe the behavior of the environment when queried by an action a_T: $\mathrm{P}[s_T \,|\, a_T, \Phi_{T-1}]$.

It is assumed in this paper that the environment has no memory, is staitonary and is purely reactive to the actions generated by the system. It is expected that the feature values returned are random with a law known a priori. An environement under hypothesis ω_k is modelled as a controlled independent identically distributed process and is completely described by a transition law $\mathrm{P}[s \,|\, a, \omega_k]$. We also restrict the spaces of features $a \in \mathcal{A}$ and feature values $s \in \mathcal{S}$ to be finite.

The complexity of the environement will be rendered by the diversity of features examined. Randomness of feature values is used as a way to model the possibly composite structure of the environment. In general, the modelling step will try to resolve a trade-off between number of features, number of hypotheses and number of available labelled data.

When assuming a controlled i.i.d. process for each hypothesis, the likelihood (2) conditionally to the hypothesis ω_k becomes:

$$\mathrm{P}_k[\Phi_T] = \mathrm{P}_k[s_T \,|\, a_T]\mathrm{P}[a_T \,|\, \Phi_{T-1}]\mathrm{P}[\Phi_{T-1}] \tag{2}$$

where for the sake of notation clarity, a probability conditioned on the hypothesis ω_k is written : $\mathrm{P}_k(\,.\,)$.

3 Stationary selective testing

This section examines the structure of the testing procedure. The central idea is the management of null likelihoods for several observations.

3.1 Likelihood-comparable observations

A basic setting for the final decision exhibiting the candidate hypothesis is a maximum likelihood test :

$$B(\Phi_T) = \arg\max_k \mathrm{P}_k(\Phi_T) \tag{3}$$

The active decision process is sequential and potentially controlled at each time. The formula (2) shows however that the only useful contribution to the computation of (3) is the product of terms $\mathrm{P}_k[s_T \,|\, a_T]$ since the choice of the action a_T sampled according to $\mathrm{P}[a_T \,|\, \Phi_{T-1}]$ is independent from any hypothesis.

We introduce the notion of "selective testing" based on the fact that a hypothesis can be discarded with infinite confidence when the corresponding conditional likelihood of an observed feature is null ($\mathrm{P}_k[s_T \,|\, a_T] == 0$). The basic idea of selective testing is to make a "comparative" test based on likelihood comparison only for the hypotheses that cannot be discarded readily.

We introduce now the following notations: $x = [s, a]$ is the composite observation state (feature value, action), S is the space of these observations and $\mathrm{S}_k = \{x \in \mathrm{S} \,/\, \mathrm{P}_k[x] > 0\}$ the support of each conditional probability where $\mathrm{P}_k[x] = \mathrm{P}_k[s \,|\, a]$.

The structure of selective testing depends on three elements: the hypotheses, the observations and the conditional likelihoods. We define a series of subsets of the observations $\mathcal{U} = \{U_p\}_{p=1\ldots|\mathcal{U}|}$ such that, in each subset U_p, the elements share the same likelihood-comparable hypotheses :

$$U_p = \{x \in \mathrm{S} \,/\, \forall k \in \Omega_p, x \in \mathrm{S}_k\} \tag{4}$$

where the coupled subsets of hypotheses Ω_p are defined as

$$\Omega_p = \{k \in \{1 \dots N\} \,/\, \forall x \in U_p, x \in \mathrm{S}_k\} \tag{5}$$

Given the conditional probability supports, there are several couples of sets $(\mathrm{S},\mathcal{U})$[1] sharing the definition above. We choose among the possible candidates the one with minimal size, which is also maximal when the sets are ordered by inclusion.

There are two consequences of this choice. The first one is that, for any subset of observations, there is a unique U_p containing all of them:

$$\forall U \subset \mathrm{S}, \exists!\, p \in \{1, \dots |\mathcal{U}|\} \text{ s.t. } U \subset U_p$$

This property implies that for any sequence Φ_T, there exists a unique element in $\mathcal{U}$ which contains all the observations $(x_1, x_2, \dots x_T)$. We note $U(\Phi_T)$ this element.

A second consequence is that the subsets U_p's define a restricted one to one mapping between 2^{S} and $2^{|\mathcal{U}|}$: $U_p \rightarrow \Omega_p$ where $p \neq p' \Rightarrow \Omega_p \neq \Omega_{p'}$. These two properties will be used by the selective test to handle the variation of active hypotheses.

3.2 Stationary sampling law

As the previous section pointed out, given a set of hypotheses and corresponding conditional probabilities, there exists two different kinds of states: selective and likelihood-comparable.

At each time T, the collected feature generated by the environment may be selective for several hypotheses. The current set of active hypotheses, i.e. that have not been discarded by previous selective states, may either remain unchanged if $x_T \in U(\Phi_{T-1})$ or reduced by the new observation.

The idea underlying a selective testing procedure is to reduce as soon as possible the number of active hypotheses by issueing actions likely to generate selective states. Indeed, the difficulty of identifying the true hypothesis depends strongly on the number of candidate possibilities. This fact will become more visible in the next section devoted to the construction of an optimal sampling strategy.

The choice of the action generated at each time should depend ideally on the whole past observed features. Because of the preceding remark, we restrict the problem of defining the sampling law to its dependency on the set of active hypotheses at each time. Given the set of current active hypotheses, the sampling law is now fully determined by conditional probabilities.

3.3 Testing procedure

The testing procedure will issue actions towards the environment sampled from the same law until no new selective observation is encountered. Each newly

[1] The mathematical objects $(\mathrm{S},\mathcal{U})$ is usually called a *hypergraph* in combinatorics [9].

observed feature value may generate a reorganization of the sampling law if it happens to be able to reject with infinite confidence one or several hypotheses.

The hypergraph $(\mathrm{S},\mathcal{U})$ depends on the set of active hypotheses. As the selective testing proceeds, elements from the set of edges $\mathcal{U}$ are desactivated each time hypotheses are removed from the list of candidates. Removing a hypothesis is equivalent to removing all its connected edges in the hypergraph on the hypotheses labels $(\{1 \dots N\}, \{\Omega_p\})$. Since there is a one to one mapping between the subsets of hypotheses labels $\{\Omega_p\}$ and the subsets of observations $\{U_p\}$, removing a hypothesis is equivalent to removing the corresponding subsets in $\mathcal{U}$. Both the set of active hypotheses and subsets of likelihood-comparable observations have to be updated when a new feature value is observed.

Given a fixed maximal number of observations $T_{\max}$, the selective testing procedure can be described the following way:

1. Update the current stationary sampling law based on the current set of active hypotheses.
2. Issue an action a_T sampled from the current stationary sampling law.
3. Collect the feature value s_T and append $[a_T, s_T]$ to the current sequence Φ_{T-1}.
4. Update the set of active hypotheses and subsets of likelihood-comparable observations.
5. If there are more than one active hypotheses or if $T < T_{\max}$ go to 1.
6. Compute the likelihoods of the remaining hypotheses.
7. The winning hypothesis is the one with highest likelihood.

The next section describes on what optimality grounds will be constructed the stationary sampling laws.

4 Optimal sampling law

The general principle for designing a sampling strategy is to find the basis for a trade-off between exploration of selective states and comparison of likelihoods. This will be achieved by computing the asymptotic rate of convergence of a selective testing given the active set of hypotheses and the conditional probabilities.

4.1 Asymptotics of selective testing with fixed sampling law

The global probability of error generated by the test (3) is defined as:

$$\mathrm{P}_e = \sum \pi_k \mathrm{P}_k[B(\Phi_T) \neq k]$$

where the π_k's are the priors.

Each term can itself be decomposed into:

$$\mathrm{P}_k[B(\Phi_T) \neq k] = \sum_{k' \neq k} \mathrm{P}_k[B(\Phi_T) = k']$$

stating that the global probability of error is a linear combination of terms of the form $\mathrm{P}_k[B(\Phi_T) = k']$.

One can decompose one step more the probability of error using the subsets of hypotheses $\mathcal{U}$ defined above. Indeed, given a sequence Φ_T, there exists a unique set $U(\Phi_T)$ containing all the observations, and we have:

$$\mathrm{P}_k[B(\Phi_T) = k'] = \sum_{p=1}^{|\mathcal{U}|} \mathrm{P}_k[B(\Phi_T) = k' \,|\, U(\Phi_T) = U_p]\, \mathrm{P}_k[U(\Phi_T) = U_p] \quad (6)$$

In this paragraph, we are interested in studying the asymptotics of the error when the observations are generated by a fixed sampling law. If $\mu(a)$ is this law, the couple $x_T = [s_T, a_T]$ becomes i.i.d. with a probability transition equal to:

$$\mathrm{P}_k[x] = \mu(a)\mathrm{P}_k(s \,|\, a) \quad (7)$$

Given the assumptions defined above (i.i.d. control process and fixed stationary sampling law) one can prove:

Proposition 1. *The probability of deciding a wrong hypothesis when the observations generated by a fixed stationary sampling law belong to a selective set U_p decreases to zero exponentially fast as the number of observations goes to infinity. The rate of convergence is defined as:*

$$\lim_{T\to\infty} -\frac{1}{T} \log P_k[B(\Phi_T) = k' \,|\, U(\Phi_T) = U_p] = \rho_p(k, k') > 0 \quad (8)$$

This result can be obtained using Large Deviation Theory [10, 11]. A first important point to note in this result is that we have an exact convergence rate, not a bound. A second point is that this rate is computable with an explicit formula. See appendix (A.1) for a proof of this result.

The second term $\mathrm{P}_k[U(\Phi_T) = U_p]$ contributing to the error is also decaying to zero exponentially fast. Indeed, it is easy to check that:

$$\lim_{T\to\infty} -\frac{1}{T} \log \mathrm{P}_k[U(\Phi_T) = U_p] = -\log \sum_{x\in U_p} \mathrm{P}_k[x] = \tau_p(k) \quad (9)$$

We therefore have the global result :

$$\lim_{T\to\infty} -\frac{1}{T} \log \mathrm{P}_k[B(\Phi_T) = k' \,|\, U(\Phi_T) = U_p]\mathrm{P}_k[U(\Phi_T) = U_p] = \rho_p(k, k') + \tau_p(k) > 0$$

stating that the probability of wrong guessing when the observations belong to the same selective set decreases to zero exponentially fast. This global rate is the sum of two terms: one qualifying the probability of staying in the same subset of observations, one quantifying the capacity of discriminating using a maximum likelihood test. The trade-off between exploration of selective states and comparison of likelihoods appears naturally in this formulation.

This last result proves that the probability of error P_e is a linear combination of terms decreasing to zero exponentially fast. It is therefore itself decaying to zero exponentially fast with a rate equal to the slowest, i.e. smallest. We have the proposition:

Proposition 2. *The probability of error of the maximum likelihood selective test decreases to zero exponentially fast when the number of observations goes to infinity with a rate equal to:*

$$\lim_{T\to\infty} -\frac{1}{T}\log P_e = \min_{k\neq k'}\min_{p}\left(\rho_p(k,k') + \tau_p(k)\right) \tag{10}$$

The rate (10) measures globally the complexity of discriminating between a set of hypotheses using a fixed sampling strategy and a selective testing procedure. It is therefore a straightforward candidate for a criterion — an energy — to optimize. The next section examines the possibility of using this criterion to find an optimal sampling strategy.

4.2 Sub-optimal fixed sampling strategy

In the process of selective sampling, the only free parameter is the sampling law μ. The probability transitions $\mathrm{P}_k(s\,|\,a)$ describe the environment and are only used in the calculation of the likelihoods.

Given the conditional probabilities, the global convergence rate (10) depends on the fixed sampling law $\mu(a)$ through (7). If we consider this rate to be a good criterion able to measure the discriminative capacity of the selective testing process, the best sampling strategy $\boldsymbol{\mu}^*$ should be defined as:

$$\min_{p,k\neq k'}\left(\rho_p(k,k';\boldsymbol{\mu}^*) + \tau_p(k;\boldsymbol{\mu}^*)\right) = \sup_{\boldsymbol{\mu}}\min_{p,k\neq k'}\left(\rho_p(k,k';\boldsymbol{\mu}) + \tau_p(k;\boldsymbol{\mu})\right) \tag{11}$$

where we have made explicit the dependency of the rates (8) and (9) on the sampling law. The supremum in (11) is actually a maximum since the rates ρ_p and τ_p are bounded and the sampling law is constrained to belong to the space of positive measures.

The optimization of (11) is a difficult problem. There is one straightforward situation — 2 hypotheses and no selective observations — for which the optimal rate is obtained when sampling the best feature: $\mu^*(a) = \mathbf{1}_{a=a^*}$ (see appendix A.2). In the general case, however, each elementary rate $\rho_p(k,k';\boldsymbol{\mu})$ is itself the result of an optimization and makes the calculation of (10) computer intensive.

The convergence of the test errors to zeros is warranted for any sampling strategy. What is sought out is a good sampling law, not necessarily uniform and querying all the features. We propose to generate a sub-optimal sampling strategy by using a linear approximation of the function $\boldsymbol{\mu} \to \rho_p(k,k';\boldsymbol{\mu})$. It is obtained by computing the rates (9) for each feature a and summing them according to:

$$\rho_p(k,k';\boldsymbol{\mu}) + \tau_p(k;\boldsymbol{\mu}) = \sum_a \mu(a).\left(\rho_p(k,k';\mathbf{1}_a) + \tau_p(k;\mathbf{1}_a)\right) \tag{12}$$

where $\mathbf{1}_a$ is the sampling law having a 1 at th a-th position and 0 elsewhere.

The linear approximation (12) makes the optimization (11) a constrained linear min-max problem which can be solved efficiently.

Due to the constraint $\sum_a \mu^*(a) = 1$, a sampling law solution will often have several null coordinates, meaning that only a few features are really useful for the discrimination of a given set of hypotheses. However, as the number of candidate hypotheses decreases due to the observation of selective feature values, the set of useful features may vary with the number of likelihood-comparable sets $\mathcal{U}$ still active at each step. This means that the "optimal" set of features depends on the nature of the recognition problem and must be adapted on line. The optimization (11) can be understood as a feature selection phase adapted to the recognition of controlled i.i.d. processes.

5 Noisy shape contour recognition

5.1 Contour detection

The application illustrating the active sampling approach is 2D shape recognition based on their noisy contour detection.

The shapes we are trying to discriminates are shown Fig. 1. They consist of 6 planes observed from above in any orientation. The typical applicative context is remote sensing from an aerial camera.

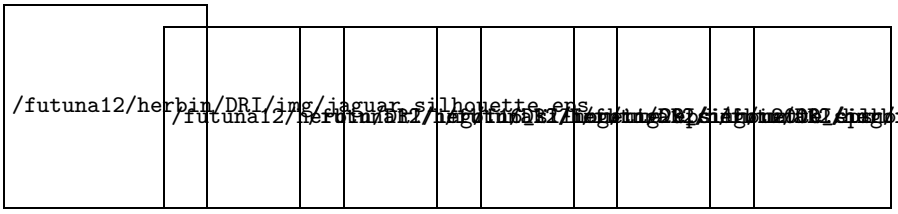

Fig. 1. The six types of shapes in random orientations. From left to right: Jaguar, F16, Mig 29, Mirage, Rafale, Sukhoi 27

It is assumed that the distance from the camera to the object is known from another source of information. Scale invariance is therefore not an issue here. The only nuisance parameters are clutter, object rotation and sub-pixel position.

Contour is a graphical primitive assumed to be quite stable to illumination variations and is often detected in a preliminary phase prior to shape representation. In practice, extracted contours are seldom closed, limiting the pertinence of many common shape representations based on their boundary [12].

Figure 2 shows some examples of contours produced by a standard algorithm (Canny) on several noisy images. The images tested were synthesized using a sensor model characterized by a gaussian transfer function, additive white noise (SNR = 15dB) and uniform random spatial sub-sampling to account for aliasing.

In the sensor model studied, shapes are between 25 to 60 pixels wide depending on the size of the plane observed and on the quality of the image feature extraction.

The contour detection process used is purposively elementary. Indeed, contours have "imperfections": areas with high local curvature are often thresholded out and sub-pixel sampling at the object boundary produces non-linear perturbations. No specific procedure to improve them by morphological closing or region growing, for instance, is attempted. Furthermore, in the application of interest, it seems illusory to base the hypothesis testing on a prior segmentation which would define the set of pixels associated with a shape: in general, the background is potentially complex and unevenly contrasted with the object. Segmentation produces often artefacts in this context difficult to neutralize. The contour imperfections generated by the detector used in this application are expected to be generic.

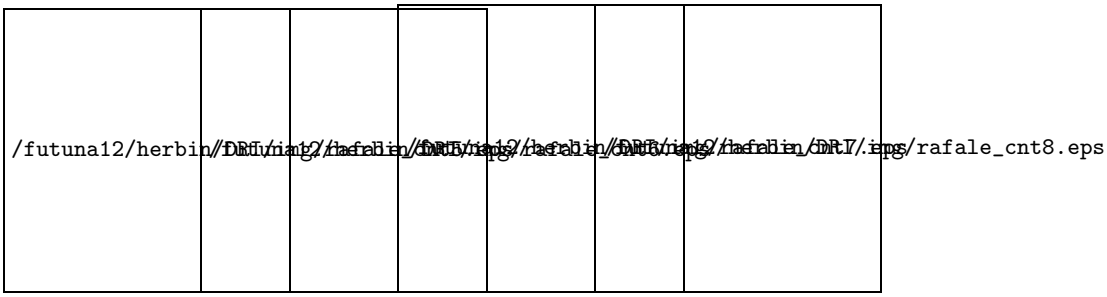

Fig. 2. Influence of noise and aliasing in contour detection. The shapes have the same orientation but various sub-pixel positions and noise outcomes.

5.2 Contour distribution coded as a controlled i.i.d. process

Contour points are local elements spatially ordered. One simple way to describe their arrangement is to exploit the distribution of specific features attached to *pairs* of points.

Bipoint feature distributions have been used in several studies to describe shapes made of random spatial arrangements of pixels or edgels. In perceptual grouping, elements are associated according to similarity, colinearity or proximity and characterized by a global grouping likelihood [13]. Random graphs between local contrast detectors, which are collections of bipoint features, can also be used for shape modelling [2].

We use bipoint feature value distributions conditioned on specific bipoint orientations to model shapes as a controlled i.i.d. processes. Bipoint orientation conditioning will be function of grey-level gradients and bipoints directions. Let $\boldsymbol{P}$ be a contour point, $\nabla \boldsymbol{P}$ the gradient computed at location $\boldsymbol{P}$ in global coordinates and Θ the function returning the direction of a vector. Gradient can be computed on the whole image using standard algorithms. A set of contour

points will be represented by several distributions indexed by the angle between local gradient and bipoint direction $\theta = \Theta(\nabla \boldsymbol{P}_1) - \Theta(\boldsymbol{P}_2 - \boldsymbol{P}_1)$ and bipoint length interval $\|\boldsymbol{P}_1 - \boldsymbol{P}_2\| \in [l_{\min}, l_{\max}]$. Define:

- the distribution of bipoint lengths: $\mathrm{P}_l\left(\|\boldsymbol{P}_1 - \boldsymbol{P}_2\| \,|\, \theta\right)$
- the distribution of gradient angles: $\mathrm{P}_g\left(\Theta(\boldsymbol{P}_1) - \Theta(\boldsymbol{P}_2) \,|\, \theta, l_{\min}, l_{\max}\right)$

The pixel size of the shapes (20 to 60) and their erratic repartition forbids the usage of discrete approximation of continuous geometric features. In this graphical context, most of the methods derive recognition from template matching [14]. These are usually not rotation invariant and require a preliminary model space compression. It is easy to check that the conditional distributions defined above are rotation invariant.

Given a set of contour points, we define a control as a direction θ and a length interval $[l_{\min}, l_{\max}]$ coupled with a type of feature — gradients angle or bipoint length. Active sampling is the operation of choosing a control and collecting the random bipoint feature value.

Figure 3 shows the set of bipoints selected by various angles θ. Different parts of the shapes are selected when the type of bipoint changes. It is expected that this way of analyzing the contour point arrangement will allow shape parts to be specifically observed when they are different while ensuring more global shape comparison for certain other bipoint distributions.

Fig. 3. Distribution of contour bipoints for various angles $\theta \in \{0, \pi/4, -\pi/4, \pi/2\}$ and bipoint length between 5 and 15 pixels.

5.3 Experiment

The shapes were modelled using two types of length intervals ($[5, 15]$ and $[10, 20]$) and 36 bipoint orientations resulting in 108 different types of queries. The gradient angles were uniformely quantized with a 5° step resulting in 72 values. The bipoint lengths were quantized with a 2 pixel length step resulting in 40 values. The controlled i.i.d. model was estimated using 200 simulated images per model.

The overall estimated model contains 10 comparative sets of hypotheses following definition (4). The corresponding sampling laws computed using the linear

approximation (12) selected between 2 and 16 queries among the 108 possible types. The maximum number of query types is associated with the comparative set containing all the hypotheses. This seems logical since the process should explore the maximum number of possible features to discriminate the maximal number of hypotheses.

Figure 4(a) shows the error decay. Two sampling strategies are compared: sub-optimal computed using (11) and uniform. The feature values are drawn from the estimated model. Sub-optimal strategy generates slightly faster error decrease, but using much fewer features.

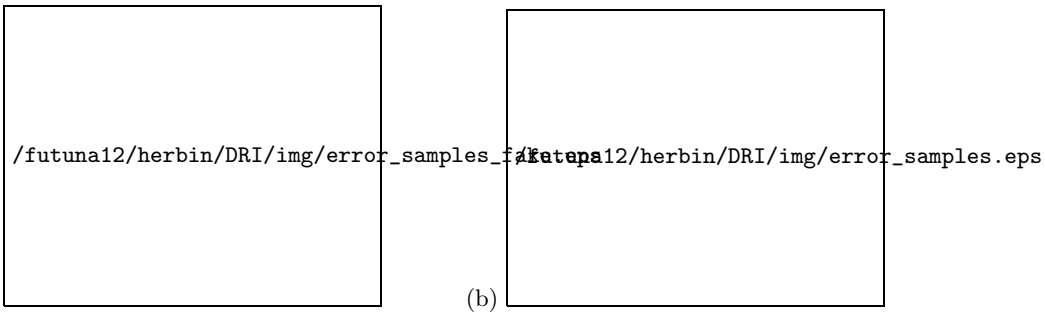

Fig. 4. Recognition error versus number of sample bipoints. (a) Empirical error using a sampling on the learned model with optimal and uniform sampling strategies. (b) Empirical error on a set of 1200 test images with optimal sampling strategy.

Selective testing was empirically evaluated on 1200 synthesized images (random orientation, SNR = 15dB, sub-pixel uniform sampling). Figure 4(b) shows the evolution of the recognition error when the number of sample bipoints increases. The performance saturates at a recognition rate of 67%.

One possible explanation of this saturation is the difference between bipoint distribution of a single shape and the mixed distribution learned from several examples. Model estimation should be able to reject several features when they appear to be badly characterized by an i.i.d. law.

The selection of optimal features concentrates the sampling process on very few sources of information: one possibility to increase robustness could be to relax the optimality criterion and add sub-optimal "good" features in the list of sampling actions.

The confusion matrix obtained after having sampled 300 bipoint feature values is shown Tab. 1. Some hypotheses are more easily recognized than others. Good recognition rates are obtained when there are only few remaining active hypotheses at the end of the active sampling process. The hypothesis su27, in particular, is generally identified by rejecting all the other candidates.

From this first experiment one should retain that model estimation must be carefully conducted. The validity of the i.i.d. assumption should be checked, and the consequences of a small number of learning examples on the selective states should be controlled. These are issues for future work.

Table 1. Confusion matrix using 300 sample bipoints by image.

	m2000	mig29	su27	jaguar	rafale	f16
m2000	**0.43**	0.09	0	0.02	0.19	0.21
mig29	0	**0.86**	0.14	0	0	0
su27	0	0	**1.00**	0	0	0
jaguar	0	0.14	0	**0.81**	0	0.05
rafale	0	0.24	0	0.04	**0.64**	0.08
f16	0.02	0.21	0	0.09	0.21	**0.47**

6 Conclusion

The specific aspects of multi-hypotheses discrimination has not been given much attention in the litterature. Global decision is often reduced to a series of binary comparisons for which, indeed, possible methods abound. The goal of this paper was to propose a general framework genuinely dedicated to the discrimination of multiple hypotheses on complex data which settles on a rigorous mathematical ground. The approach was demonstrated on a 2D noisy shape recognition problem and shows promising preliminary results.

The coupling of optimal active sampling and selective testing can be understood as unifying two different questions: *feature selection* since the optimal sampling strategy selects the best queries according to a multi-hypothesis discrimination criterion, and *data fusion* since the final test gathers the feature values collected in a global maximum likelihood decision.

The basic setting developped in this article can be improved in several ways. A notion of memory, e.g. a Markov dependence or a short-term buffer, may be introduced in both the sampling law and the environment modelling in order to design more flexible strategies. The accuracy of model estimation is very influential and necessitates further studies to control its impact on the decision process.

A Calculation of the error rate of convergence

A.1 Maximum likelihood test

The rates of convergence can be computed exactly thanks to tools developped in the Large Deviation Theory of empirical processes. This section is devoted to presenting the basic useful results. Refer to [11, 10] for a more complete presentation.

Define a rate function as:

$$I(\boldsymbol{y}) = \sup_{\boldsymbol{\theta}>0} \left[\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle - \log M(\boldsymbol{\theta})\right] \tag{13}$$

The Laplace transform $M_k(\boldsymbol{\theta})$ is defined as:

$$M(\boldsymbol{\theta}) = \mathrm{E}\left[\exp\langle\boldsymbol{\theta}, \boldsymbol{l}(x)\rangle\right] \tag{14}$$

where $\boldsymbol{l}(x)$ is a vector valued function in $\mathbb{R}^k$ of the observations x, and the expectation is computed on their distribution.

The fundamental result (Cramér Theorem) characterizes the occurence of rare deviations from the empirical mean:

$$L_T = \frac{1}{T}\sum_{t=1}^{T} l(x_t)$$

Theorem 1. *For any set $A \subset \mathbb{R}^d$, we have*

$$\liminf_{T->\infty} -\frac{1}{T}\log P\{L_T \in A\} \geq \inf_{\boldsymbol{y}\in\bar{A}} I(\boldsymbol{y})$$
$$\limsup_{T->\infty} -\frac{1}{T}\log P\{L_T \in A\} \leq \inf_{\boldsymbol{y}\in A^\circ} I(\boldsymbol{y})$$

If the two bounds coincide, which is the case in the application we are studying, we have:

$$\lim_{T->\infty} -\frac{1}{T}\log \mathrm{P}\{L_T \in A\} = \inf_{\boldsymbol{y}\in A} I(\boldsymbol{y})$$

In the multihypothesis likelihood test, we are studying the random behavior of the log-likelihood vector in $\mathbb{R}^N$ defined by:

$$\boldsymbol{l}(x) = \log\left(\mathrm{P}_1[x], \ldots \mathrm{P}_N[x]\right)$$

when the samples are drawn from hypothesis k.

As a direct consequence of the above theorem, we have:

Proposition 3. *The error rate of the probability of deciding a wrong hypothesis k' is:*

$$\rho(k, k') = \inf_{\boldsymbol{y}\in C(k')} I_k(\boldsymbol{y}) \tag{15}$$

where the convex constraint is defined as:

$$C(k') = \{\boldsymbol{y} \,/\, \forall j \neq k', y_{k'} > y_j\}$$

The practical computation of (15) is managed using elementary convex analysis. Indeed, one can prove:

Theorem 2. *The rate of convergence (15) can be computed as:*

$$\rho(k,k') = -\inf_{\boldsymbol{\lambda}>0} \log M_k(-\tilde{\boldsymbol{\lambda}}_{k'}) \tag{16}$$

where $\boldsymbol{\lambda}$ is the vector of $N-1$ Lagrange multipliers, and $\boldsymbol{\lambda}_{k'}$ is the vector of Lagrange multipliers augmented by a normalizing factor at the k'-th position:

$$\tilde{\boldsymbol{\lambda}}_{k'} = (\lambda_1 \dots \lambda_{k'-1}, -\sum_j \lambda_j, \lambda_{k'+1} \dots \lambda_{N-1}).$$

The Laplace transform based on the Lagrange multipliers is now defined as:

$$M_k(-\tilde{\boldsymbol{\lambda}}_{k'}) = \sum_x \mathrm{P}_k[x] \prod_{l \neq k'} \left(\frac{\mathrm{P}_{k'}[x]}{\mathrm{P}_l[x]} \right)^{\lambda_l} \tag{17}$$

The numerical value of the rate is obtained using optimization (16).

A.2 Mixing active sampling and selective testing

The results from the previous section assumed that the log-likelihood vector was defined for any observation. The principle of selective testing is to exploit the probability that some of its coordinates become infinite. The active sampling principle consists in randomly examining a family of random laws.

We assume now that the state x is composite $x = [s, a]$ where s is assumed to be a finite value, and a a given action governed by a sampling strategy $\boldsymbol{\mu}$:

$$\mathrm{P}_k[x] = \mathrm{P}_k[s,a] = \mu(a)\mathrm{P}_k(s \,|\, a)$$

Define a likelihood-comparable set U_p of states, and $U_p(a)$ the feature values which are comparable for a given sampling a action:

$$U_p(a) = \{ s \,/\, [s,a] \in U_p \}$$

and the normalized conditional probability as:

$$\tilde{\mathrm{P}}_k(s \,|\, a) = \frac{\mathrm{P}_k(s \,|\, a)}{\sum_{a'} \mu(a') \sum_{s \in U_p(a')} \mathrm{P}_k(s \,|\, a')} \tag{18}$$

In the calculation of the error rate $\mathrm{P}_k[B(\Phi_T) = k' \,|\, U(\Phi_T) = U_p]$, the only change appears in the definition of the Laplace transform (17):

$$\tilde{M}_k(-\boldsymbol{\lambda}_{k'}) = \sum_a \mu(a) \sum_{s \in U_p(a)} \tilde{\mathrm{P}}_k(s \,|\, a) \prod_{l \neq k'} \left(\frac{\mathrm{P}_{k'}[x]}{\mathrm{P}_l[x]} \right)^{\lambda_l} \tag{19}$$

where the expectation is now taken over the likelihood-comparable observations. The optimization (16) remains unchanged and we have:

$$\rho_p(k,k') = -\inf_{\boldsymbol{\lambda}>0} \log \tilde{M}_k(-\tilde{\boldsymbol{\lambda}}_{k'})$$

The dependence of the rate $\rho(k,k')$ on the sampling law $\boldsymbol{\mu}$ is strongly influenced by the presence of selective observations. If all the observations are likelihood-comparable, the rate function (13) has interesting properties [15]:

– the Laplace transform (14) is concave in $\boldsymbol{\mu}$;
– the rate function (13) is jointly convex in $\boldsymbol{y}$ and $\boldsymbol{\mu}$;
– the error rate $\rho(k, k')$ is convex in $\boldsymbol{\mu}$.

These properties make the two-hypothesis case search for optimal decay rate easy: the best achievable rate is obtained at the vertices of the sampling law. In the multi-hypothesis case, however, the search for the best achievable rate appears to be a non convex problem.

When there is a chance to observe selective observations, the convexity properties of the rate function are no longer true due to the normalizing (18). This makes the search for an optimal sampling law even more difficult, and justifies the use of a linear approximation in the computations.

References

1. Breiman, L., Stone, C., Olshen, R., Friedman, J.: Classification and Regression Trees. Wadsworth (1984)
2. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Computation **9** (1997) 1545–1588
3. Azencott, R., Vayatis, N.: Refined exponential rates in Vapnik-Chervonenkis inequalities. C. R. Acad. Sci., Paris, Math., Ser. I **332** (2001) 563–568
4. Yuille, A., Coughlan, J., Wu, Y., Zhu, S.: Order parameters for detecting target curves in images: When does high level knowledge help? International Journal of Computer Vision **41** (2001) 9–33
5. Wu, Y., Zhu, S., Liu, X.: Equivalence of julesz ensembles and FRAME models. International Journal of Computer Vision **38** (2000) 245–261
6. Herbin, S.: Recognizing 3D objects by generating random actions. In: IEEE Conference on Computer Vision and Pattern Recognition. (1996) 35–40
7. Herbin, S.: Combining geometric and probabilistic structure for active recognition of 3D objects. In: European Conference on Computer Vision. Volume 1407 of Lecture Notes in Computer Science., Berlin, Springer Verlag (1998) 748–764
8. Herbin, S.: Similarity measures between feature maps - application to texture comparison. In: Proceedings of the Texture 2002 workshop. (2002) 67–72
9. Berge, C.: Graphes et hypergraphes. Dunod (1970)
10. Kazakos, D.: Asymptotic error probability expressions for multihypothesis testing using multisensor data. IEEE Trans. Systems, Man and Cybernetics **21** (1991) 1101–1114
11. Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications. Jones and Bartlett Publishers, Boston (1993)
12. Loncaric, S.: A survey of shape analysis techniques. Pattern Recognition **31** (1998) 983–1001
13. Amir, A., Lindenbaum, M.: A generic grouping algorithm and its quantitative analysis. IEEE Trans. Pattern Analysis and Machine Intelligence **20** (1998) 168–185
14. Olson, C., Huttenlocher, D.: Automatic target recognition by matching oriented edge pixels. IEEE Trans. Image Processing **6** (1997) 103–113
15. Shimkin, N.: Extremal large deviations in controlled i.i.d. processes with applications to hypothesis testing. Adv. Appl. Probab. **25** (1993) 875–894