Recognizing 3D Objects by Generating Random Actions

Stéphane HERBIN

Ecole Normale Supérieure de Cachan
Centre de Mathématiques et de Leurs Applications / Groupe DIAM
CNRS, URA 1611
61, Av. Président Wilson
94235 Cachan Cedex, France

Abstract

This paper presents a formal model of an active recognition system that can be programmed by learning. At each time step the system decides between producing an action to generate new data and stopping to issue the name of the object observed. The actions can be directed either towards the external environment or towards the internal perceptual system of the agent. The decision strategy is based on a quantitative evaluation of the system learning experience.

The problem studied is the recognition of chess pieces using a moving camera and a multiscale feature detector. The recognition is difficult because the objects are complex – neither polyhedral nor smooth – and rather similar between classes, especially in certain view configurations. The system uses the information obtained by observing internal state transitions when the camera is moved or when the feature detector scale is changed. A simulation of the agent and the environment is used for experimental measures of the model performances.

1 Introduction

The research in tri-dimensional object recognition aims at incorporating into a general pattern recognition framework some particularities of the 3D world. The goal of pattern recognition can be seen as a meaning construction and is organized as a bottom-up and data-driven sequence of static information transformations, starting from the rough signals to the meaningful classes or names. It consists usually of four stages: segmentation, feature extraction, representation or feature selection and classification.

Most of the research in computer vision has focussed on the problem of finding the most acurate and general way of representing data in an object model. It is usually admitted that the better the representation, the easier the recognition.

The search for a universal and optimal representation has put aside the study of the recognition for itself. However, all the different stages do cooperate to produce a name or an object class, and depend strongly on the context of use. The emphasis on the representational paradigm may be misleading, and the universality of an object recognizer looked for at the wrong place.

The specificity of 3D object recognition modelling compared with standard pattern recognition is to be confronted with two difficult questions: a recognition system can not have access to or gather as a whole all the useful data and a recognition process is only valid for a given restricted context of observation. The usual pattern recognition approach which consists in developping complex object models able to summarize all what can be known about an object, independently of their future usage, and then producing a possibly costly indexing or matching search is often stuck with incomplete data management, weak robustness and combinatorial explosion. It is proposed to get round these problems by allowing the recognition system to be both *specialized* and *dynamic*.

The dynamical aspect of a recognition system requires an *active* data management. The recognition specialization will be obtained through *learning* which can be viewed as a contextualization of the process parameters.

This paper presents a formal model of an active recognition system that can be programmed by learning. At each time step the system decides between producing an action to generate new data and stopping to issue the name of the object observed. The actions can be directed either towards the external environment or towards the internal perceptual system of the agent. The decision strategy is based on a quantitative evaluation of the system learning experience.

2 Related work

The main points of the model presented in this paper is the combination of learning and active management of data without any explicit representation of the object type. These aspects are usually absent in traditional models of 3D object recognition [1] where the emphasis is on the problem of 3D information representation and matching rather than on complex recognition strategies.

The dynamic aspects of data management are well represented in the field of active vision [2, 3, 4] but do not deal with learning paradigms.

Neural networks domain is especially interested in designing learning algorithms but the research emphasis has been mostly on low level vision, attentional mechanisms and static recognition studies. The only works that integrate in the same framework learning, 3D vision and dynamic evolution are [5, 6]. They do not present however any clear evaluation of the recognition performances.

Sequential decision problems are an old question in statistical literature [7, 8]. However most of the tractable results concern either independent identically distributed random variables, perfectly estimated probabilistic environments or asymptotic behaviors and can not be directly applied in active recognition problems.

3 Model of an active recognition system

RECOGNITION AS AN ACTION

The recognition skill will be considered embedded in an *autonomous agent* interacting with an unknown, partially observable *environment*. A mobile robot with a camera would be a good example of such a system. The agent is aware of several potential names to be associated with its experience, has limited but controllable perceptual capacities, and may require the assistance of a teacher who knows the object names.

The agent is able to produce *external* actions towards its environment and *internal* actions in direction of its own perceptual system. The action of naming the object can be both internal and external since it influences the agent itself by stopping the recognition and the environment by issuing a name. This space of terminal actions will be denoted by $\mathcal{A}^{\odot}$ and constitutes the set of final decisions.

At each instant time t, the agent has an internal state s_t and can produce an action a_t which can be defined as the values taken by the random variables S_t and A_t. The goal of the agent is to guess the class $w^{\odot} \in \mathcal{W}^{\odot}$ of an observed object.

The whole internal past can be collected at time t in an information vector $\phi_t = (s_0, s_1, \ldots s_{t+1})$, which will be considered as the value taken by a random variable Φ_t. This vector summarizes all the data the system will be able to use to control its recognition process.

The evolution of the agent internal states will be modelled by a transition probability:

$$\mathbf{P}[S_{t+1} = s_{t+1} \,|\, \Phi_t = \phi_t, A_t = a_t, W^{\odot} = w^{\odot}] \qquad (1)$$

This law is the probability that the internal state is s_{t+1} at time $t+1$ when the sequence ϕ_t has been recorded, the action a_t has been produced and the object $w^{\odot}$ is being observed. This equation summarizes two different characteristics of the agent: its perceptual ability and memorizing capacity. Both can be influenced by an action. In particular, the set of observed features that characterize the object can be dynamically modified during the recognition process in order to produce multiscale analysis.

The purpose of active recognition is to design a rule that associates an action a_t to a sequence of past internal states ϕ_t for each time t. At each time, the agent will have to decide between generating a new internal state by acting, producing the guessed name of the object or giving up. The process stops when a terminal action is produced.

It will be assumed that the recognition system is dealing with a finite number N of object types $\{w_1^{\odot}, w_2^{\odot} \ldots w_N^{\odot}\}$. The terminal actions, the namings, will be also in finite number and will be symbolized by $\{a_1^{\odot}, a_2^{\odot} \ldots a_N^{\odot}\}$. It may happen that the system is unable to produce a name and stays undecided: this situation will be represented by a terminal action $a_0^{\odot}$ meaning that the system renounces issuing new actions and gives up. The space $\mathcal{A}^{\odot}$ of terminal actions, thus, consists of $N+1$ actions $\{a_l^{\odot}\}_{l \in \{0,1,\ldots N\}}$. The other actions, however, may be of any type, but it will be assumed also in the following that they are in finite number.

CONTROL LAWS

The fundamental principle of an active recognition system is to delay the final decisive action in order to collect new informative data. The role of the control law at any time is to choose the available actions and to elect among them the most appropriate one. One important conflict to solve is to decide to issue a terminal action or not. Thus, an active recognition system needs the definition of three different control features:

- A stopping rule: σ.
- A law for non terminal action generation: μ.
- A final decision rule: π.

All these laws, deterministic or stochastic, will depend on the system experience, *i.e.* the sequence vector ϕ_t. The main problem with this parameter is its growing dimension with t which makes difficult the programmation and parametrization of the control laws. One way to overcome this problem is to consider the *a posteriori probabilities* $\hat{\mathbf{w}}^{\odot}(\phi) = (\mathbf{P}[w_1^{\odot} \,|\, \phi], \mathbf{P}[w_2^{\odot} \,|\, \phi], \ldots, \mathbf{P}[w_N^{\odot} \,|\, \phi])$ as a reasonable summary of the useful information contained in ϕ_t.

The non terminal action law μ will be purely random and will not depend on the past. At each time, if it is decided not to stop, a random action will be chosen according to the probability $\mu(a)$.

The final decision π can be considered as a Bayes decision rule, where the choice of the object name is delayed until a stopping signal is issued: it will be based also on the a posteriori probabilities $\mathbf{P}[w^{\odot} \,|\, \phi_t]$, and on a *rejection threshold* λ [9], if the information provided by the environment is not discriminative enough.

The system will continue generating actively new data until the a posteriori probabilities become discriminating enough. The stopping rule σ should base its decision on some global measure of the distribution such as the entropy or the probability of the most probable object. The system may decide also to limit the number of actions generated and issue a stopping signal after a fixed time threshold. The stopping rule therefore will depend on the current time and on the a posteriori probabilities.

COMPUTATION OF THE A POSTERIORI PROBABILITIES

When the pattern ϕ is a growing sequence of observations with an ever increasing size, estimating the priors and class-conditional probabilities for all possible $\{\phi_t\}_{t\geq 0}$ is often an intractable problem.

The generation of a new internal state s_{t+1} is the result of an action a_t. Every time a new state is observed, the a posteriori probabilities are modified so that the uncertainty measure on the type of object observed is refined. Indeed, the a posteriori probabilities follow the relation:

$$\mathbf{P}[w^{\odot} \mid \phi_{t+1}] = \frac{\mathbf{P}[s_{t+1} \mid \phi_t, w^{\odot}]}{\mathbf{P}[s_{t+1} \mid \phi_t]} \mathbf{P}[w^{\odot} \mid \phi_t] \qquad (2)$$

obtained as a consequence of the Bayes law.

The formula (2) gives a iterative method for computing the a posteriori probabilities when the series of laws $\mathbf{P}[s_{t+1} \mid \phi_t, w^{\odot}]$ is known. The problem of estimating and storing those probabilities is intractable in practice, since it is not known in advance what should be the useful length of the information vector ϕ_t. In order to reduce the amount of storage requirement, it will be assumed that the internal states are well described by a homogeneous Markov chain.

With this new hypothesis the formula (2) reduces to:

$$\mathbf{P}[w^{\odot} \mid \phi_{t+1}] = \frac{\mathbf{P}[s_{t+1} \mid s_t, w^{\odot}]}{\mathbf{P}[s_{t+1} \mid s_t]} \mathbf{P}[w^{\odot} \mid \phi_t] \qquad (3)$$

which makes the iterative computation of the a posteriori probabilities depend only on the transition probabilities $\mathbf{P}[s' \mid s, w^{\odot}]$ and the priors $\mathbf{P}[w^{\odot}]$ and $\mathbf{P}[s]$.

PROBABILITY ESTIMATION

The iterative computation of (3) requires the estimation of the class-conditional transition probabilities: $p_{ij}(k) = \mathbf{P}[S_{t+1} = i \mid S_t = j, W^{\odot} = w_k^{\odot}]$ and the Markov chain coefficients: $p_{ij} = \mathbf{P}[S_{t+1} = i \mid S_t = j]$. This estimation can be realized using classical maximum likelihood estimators which consist in observing the frequencies of occurence of each state on several trajectories.

The estimation for all the transition probabilities do not have the same quality. It can not be ensured that all states will be visited with the same frequency due to the Markovian structure. The estimation of the $p_{ij}(k)$'s, in this case, may be biased, and the errors may cumulate in the iterative computation of the a posteriori probabilities (3).

The a posteriori probabilities updating should therefore be conducted carefully so as to prevent the estimation error cumulation in the iterative computations. For any observation, it should be decided whether it brings to the system reliable information, *i.e.* whether it contains some safe discriminative power, given the system learning experience. This can be measured by the computation of *confidence intervals* for each estimation.

For a fixed *confidence level* $1-2\alpha$, the information provided by the observation of a transition $s_t \rightarrow s_{t+1}$ will be considered reliable if there is an object type for which the values contained in its confidence interval are clearly bigger than the values contained in the confidence intervals of the other object types. When the level $1-2\alpha$ increases, the confidence intervals get wider and the acceptance criterion becomes harder to satisfy.

If the transition observed is declared reliably informative, the iteration (3) is performed, ensuring that the a posteriori probabilities are modified in a very conservative way since big variations are controlled.

SUMMARY OF THE MODEL

With the above simplifying hypotheses, we are now able to propose a simple model of an active recognition system:

- A finite set of internal states labelled by $\{1, 2, ..., M\}$
- A finite set of terminal actions $\{a_l^{\odot}\}_{l\in\{0,1,...,N\}}$
- A finite set of names or object types $\{w_k^{\odot}\}_{k\in\{1,...,N\}}$
- Transition probabilities between states $p_{ij}(k) = \mathbf{P}[S_{t+1} = i \mid S_t = j, W^{\odot} = w_k^{\odot}]$
- A confidence parameter α for accepting the information provided by an internal state transition.
- A stationnary probability law for non terminal action generation $\mu(a)$
- A Bayes decision rule $\pi(\hat{\mathbf{w}}^{\odot}(\phi))$ with rejection threshold λ for issuing a terminal action.
- A rule for stopping the system $\sigma(\hat{\mathbf{w}}^{\odot}(\phi), t)$ based on some certainty threshold and the maximal number of actions allowed in a recognition phase, $T_{\max}$.

This model will be assumed in the following. The next section presents the simulation of a 3D recognition system where some relations between the above parameters will be experimentally studied.

4 Simulation of a 3D recognition system

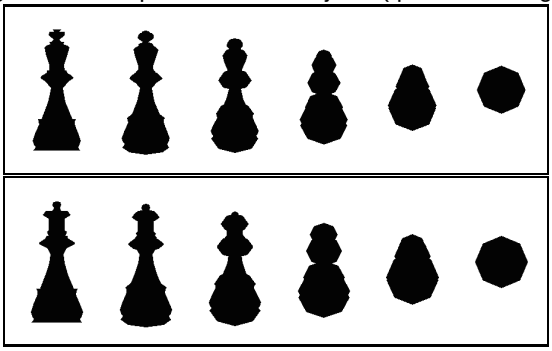
The active recognition model will be applied to a simple simulated tri-dimensional environment. The available information is in the form of segmented images, where the object pattern has been separated from the background. The system is able to extract some features from the object pattern at different scales and can generate new images by changing its point of view. The sequence of image features will be characteristic of the object observed (Fig. 1).

4.1 Multiscale feature detection

The objects observed project onto the camera retina as a connex area delimited by a piecewise smooth curve: the image outline.

This outline is itself the partial projection of the occluding contour, which consists of the set of points on the object where the light of sight is tangent to the object and of the curves defining the intersection of the differentiable surfaces.

Fig. 1: View sequences of two objects (queen and king).



The organization of the outline singular points, and its variation when the view point is continuously modified, follow specific rules: the configuration of of points — the *aspect* — is topologically stable, except when the view point crosses specific hypersurfaces, *i.e.* when a visual event occurs [10].

The organization of the visual events may be characterized by an *aspect graph* [11] which is a differential structure summarizing the possible variations of the outline singularity topology. Its exact computation requires many calculations and has only been realized rigorously for algebraic surfaces [12]. In the case of complex objects it is untractable in practice. The main result about the theory of singularities that will appear useful here is that objects can be characterized by the evolution of the outline singularities when the direction of observation is continuously modified.

The concept of singularity in a discrete space is ill defined. Instead, we substitute to it the idea of extremal variation of the contour tangent direction.

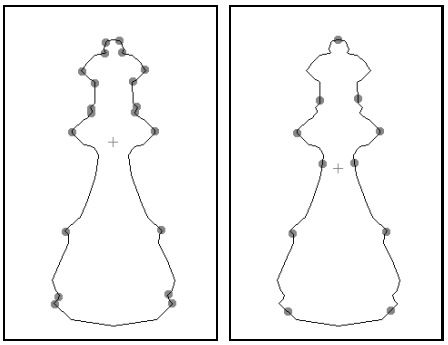


Fig. 2: Singular point detection for two different scales.

We extract first the outline by contour-following. The rough 1D signal thus obtained will be filtered in two phases: first by convoluting with a gaussian kernel and then by eliminating the local extrema under a given threshold. The first operation allows one to introduce a scale factor (the width of the gaussian kernel) and a detection threshold. This singular point detection may be regarded as an adaptation of wavelet based filters [13] and is simple to implement. Fig. 2 shows two singular point detections for two different gaussian kernels. We divide finally the one-dimensional chain into a fixed number of windows and pick out the extremum with highest absolute value as a representant. This sampling operation produces fixed length vectors with positive value if the extremum is a convex point, negative value if it is a concave point and zero if no extremum over the detection threshold has been found.

The concept of aspect change will also necessitate an adaptation. A topological aspect is a class of outlines that share the same singularity structure. By extension, a *generalized aspect* can be defined as a class of curves that share some similarity.

The curves, as described above, are represented by fixed length vectors, where each coordinate indicates if a singular point has been found in a given area. The usual euclidean distance between those vectors can be considered as a meaningful measure of similarity between curves since the maximal dissimilarity between curves occurs when the singular points shift from concave to convex and vice versa.

Once a meaningful similarity measure has been found in a vector space, standard clustering or vector quantization algorithms can be used to define the generalized aspects. We chose in this simulation the "Neural Gas" [14], which is a competitive learning algorithm that showed experimentally its ability to represent with low compression error complex and polymodal distributions. An aspect will be represented by one of the prototype vector.

The recognition system will base its decisions on the generalized aspects and on their evolution when actions are produced. The internal states will be taken as labels of the generalized aspects defined by the vector quantization prototypes.

4.2 Modular organization

The active recognition model will be realized in a simulated environment which makes easier the study of the parameters influence. Each of the modules needed for the simulation were integrated in the environment provided by the software AVS (Advanced Visual System) which contains several facilities for visualizing data, controlling data flow and already includes predefined graphic routines.

The active recognition system simulation will be divided into several modules (Fig. 3):

- a feature extractor and vector quantizer, taking for input an image and issuing as output a generalized aspect label. It can be parametrized by a detection scale.
- a decision module selecting the actions on the environment and on the feature extractor. Its input is an aspect and its output is an action label.
- a series of modules instantiating the actions chosen by the decision module.
- an environment consisting of a camera and an object. Its input is a camera position and its output is the image of the camera retina.

In the simulation tested, the possible actions were of three types: a switch between two different feature detection scales, an incremental move of the camera upward with a constant angular step, an incremental move downward. The camera needs only one degree of freedom since we are using object with a symmetry of revolution.

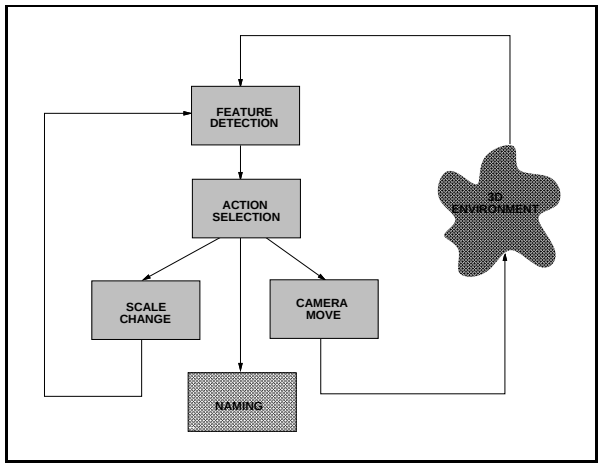


Fig. 3: General 3D active recognition scheme.

4.3 Experimental results

The performance of the recognition system was measured on the problem of differentiating between two types of objects, kings and queens, each class containing 6 different objects generated by random deformation of the same polyhedral CAD model.

The camera was able to incrementally move around the observed object with a $5°$ increment in a vertical plane.

The singular points configurations were extracted using two scales and condensed into 20 possible generalized aspects by the vector quantizer algorithm, making the number of transition probabilities to learn equal to $20 \times 20 \times 2 = 800$.

The coefficients of the Markov chain $p_{ij}(k)$ were estimated using 10000 observed random transitions where the name of the object was known. Many of the possible transitions between aspects were not observed which should not be surprising since in general aspect graphs are poorly connected.

The system stops and issues a name when the a posteriori probabilities $\hat{\mathbf{w}}^{\odot}(\phi_t)$ become discriminating enough or when the recognition time is elapsed. The a posteriori probabilities are declared discriminating enough when $\max_k \mathbf{P}[w_k^{\odot} \mid \phi_t]$ is above a given threshold taken equal to the Bayes rejection threshold λ. Therefore the system continues issueing uniformly distributed random actions until the Bayes threshold is reached, unless it is forced to stop because of a time limit $T_{\max}$.

The behaviour of the system depends thus on several parameters: the rejection threshold λ, the confidence parameter $1 - 2\alpha$ and the maximal terminal time $T_{\max}$.

Fig. 4 shows the recognition performance for a fixed confidence level ($\alpha = 1\%$). Each point was obtained by averaging over 5000 experiments. Those measures show that, for a given rejection threshold, both erroneous and correct recognition ratios increase with the time limit $T_{\max}$. This means that, if we let the system generate new data until its a posteriori probability reaches a fixed discriminating threshold, the recognition error can not go below a certain bound: the overall recognition capacity of the system is limited.

However the correct recognition ratio increases when the rejection threshold decreases if there is no time limit ($T_{\max} = \infty$), and does not seem to be bounded by an upper limit. This suggests that any performance quality can be expected if the system is able to generate any number of actions.

Fig. 4 shows also the difference between active recognition and the static Bayes decision obtained for $T_{\max} = 0$: the recognition error is usually lower for the Bayes law, but is obtained at the expense of a high rejection ratio and low recognition ability.

The average terminal time, as should be reasonably expected, decreases when the rejection threshold (Fig. 6) increases or when the confidence parameter (Fig. 4) increases: it takes longer to produce reliable and discriminative information. Logically also, the correct recognition ratio increases with the confidence parameter (Fig. 4).

Fig. 7 shows a diagram characterizing the overall performance of an active recognition system. The specific behaviour of an unbounded sequence of actions ($T_{\max} = \infty$) is clearly noticeable: when the number of actions is not limited, the correct recognition ratio *and* the erroneous recognition ratio can be improved by decreasing the rejection threshold.

5 Conclusion

The formalism presented in this paper aimed at showing that an alternative to the object model paradigm is possible. An accurate recognition can be achieved without any explicite representation of the environment. The statistical framework provides many mathematical tools for the performance analysis and the control of a recognition strategy.

The active recognition model can be improved in several ways. The sequence of actions is informative also and can influence the decision processes. The actions were generated randomly and did not use any information provided by the past ϕ_t. The action law μ could therefore be designed to search more directly discriminative data based on the immediate experience.

Acknowledgements

The author was supported by a scolarship from the Centre National de la Recherche Scientifique. This paper describes part of a Ph.D work directed by Robert Azencott.

Fig. 4: Erroneous and correct recognition ratios in function of rejection threshold for various $T_{\max}$'s.

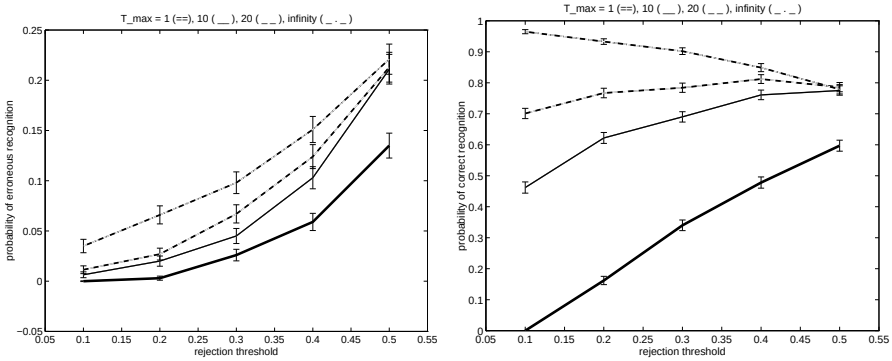


Fig. 5: Correct recognition ratio and average trajectory length for various acceptance levels.

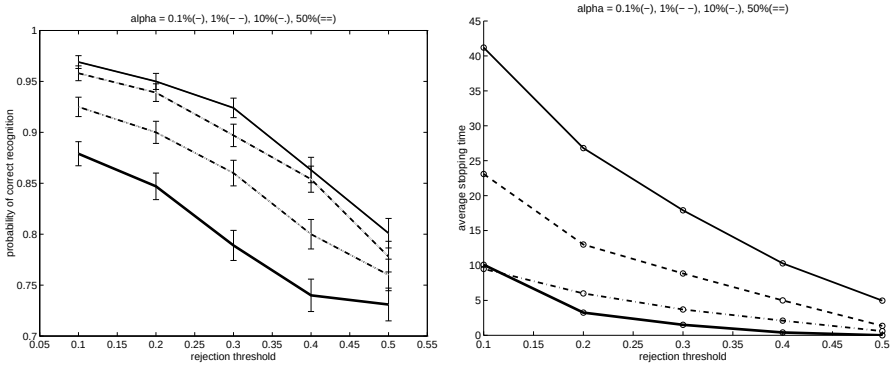


Fig. 6: Average recognition trajectory length in function of rejection threshold.

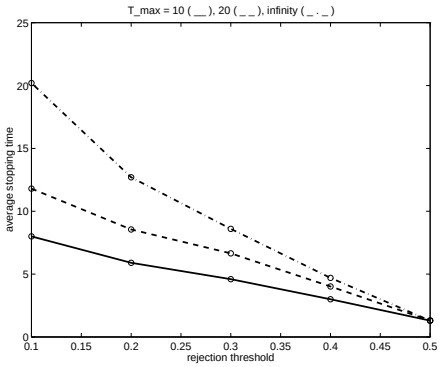
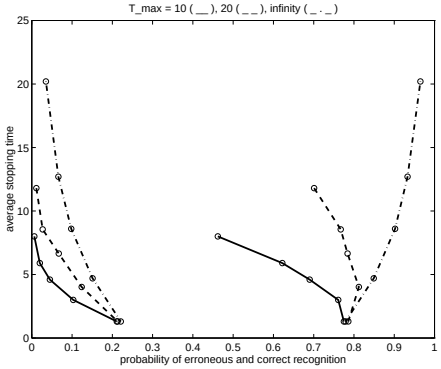


Fig. 7: Average recognition trajectory length in function of erroneous and correct recognition ratios.



References

[1] A. Jain and P. Flynn, eds., *Three-Dimensional Object Recognition Systems*, vol. **1** of *Advances in image communication*. Amsterdam: Elsevier, 1993.

[2] R. Rimey and C. Brown, "Task-oriented vision with multiple bayes nets," in *Active Vision* (A. Blake and A. Yuille, eds.), ch. 14, pp. 217–236, Cambridge Ma: MIT Press, 1992.

[3] S. Dickinson, H. Christensen, J. Tsotos, and G. Olofsson, "Active object recognition integrating attention and viewpoint control," in *Computer Vision – ECCV'94* (J.-O. Eklundh, ed.), no. 801 in Lecture Notes in Computer Science, pp. 3–14, Berlin: Springer Verlag, 1994.

[4] S. Vinther and R. Cipolla, "Active 3d object recognition using 3d affine invariants," in *Computer Vision – ECCV'94* (J.-O. Eklundh, ed.), no. 801 in Lecture Notes in Computer Science, pp. 15–24, Berlin: Springer Verlag, 1994.

[5] M. Seibert and A. Waxman, "Adapative 3-D-object recognition from multiple views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. **11**, pp. 107–124, 1992.

[6] G. Bradski and S. Grossberg, "Fast learning VIEWNET architectures for recognizing three-dimensional objects from multiple two-dimensional views," *Neural Networks*, vol. **8**, no. 7/8, pp. 1053–1080, 1995.

[7] A. Wald, *Sequential Analysis*. New York: John Wiley, 1947.

[8] T. Ferguson, *Mathematical Statistics: a decision theoretic approach*. New york: Academic Press, 1967.

[9] P. Devijver and J. Kittler, *Pattern Recognition: a Statistical Approach*. London: Prentice Hall, 1982.

[10] J. Rieger, "On the classification of views of piecewise-smooth objects," *Image and vision computing*, vol. **5**, pp. 91–97, 1987.

[11] J. Koenderink, *Solid Shape*. Cambridge Ma: MIT Press, 1990.

[12] S. Petitjean, J. Ponce, and D. Kriegman, "Computing exact aspect graphs of curved objects: algebraic surfaces," *International Journal of Computer Vision*, vol. **9**, no. 3, pp. 231–255, 1992.

[13] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. **8**, pp. 679–698, 1986.

[14] T. Martinetz and K. Schulten, "Topology representing networks," *Neural Networks*, vol. **7**, no. 3, pp. 507–522, 1994.