F. Solina
W. G. Kropatsch
R. Klette
R. Bajcsy (eds.)

Advances in Computer Vision

Advances in Computing Science

SpringerWienNewYork

F. Solina
W. G. Kropatsch
R. Klette
R. Bajcsy (eds.)

Advances in Computer Vision

SpringerWienNewYork

Contents

Optimal models for visual recognition [1]

Matevž Kovačič, Bojan Kverh, and Franc Solina

1 Introduction

Over the years building models of objects from sensory data has been tackled in various ways. Following [1], model based recognition methods are divided into graph theoretic and non graph theoretic. Graph theoretic methods use graphs as a representation for objects and scenes. An object is divided into parts. Nodes of a graph that describes an object characterize the parts of the object and arcs of the graph represent spatial relations among parts of the object. Recognition of an object in the scene is performed as search for a subgraph isomorphism between the scene graph and each of the model graphs. In non graph theoretic methods, local features are used to describe the object. Grimson and Lozano-Peres [3], used a constrained tree search to efficiently coordinate values of point features and surface normals in models to those found in the scenes.
We are not interested in comparing the efficiency of the models in terms of time and space complexity, but to select the most *probable* model among the predetermined set of classes of models.
Suppose the scene consists of an object represented as a set of points in a two dimensional plane. For the sake of the argument, let us limit the set of possible models to the set of single valued functions. If there are n such points then the object can always be explained by a model of a form of a polynomial of degree up to $n-1$. Statistical measures would prefer such a model over, say, a much simpler model such as a linear curve which, for example, misclassifies a single instance in our set of observations. Clearly, the measure of a model justification must take into account the complexity and the accuracy of a model.
In general, there is an infinite number of models which explain the data. A crucial question is which model to choose. There are four principles for model justification. William of Ockham proposed the principle known as *Occam's razor:* if there are alternative explanations for a phenomenon, then, all other things being equal, we should select the simplest one[2]. Identification of the 'simplicity of an object' with 'an object having short effective description' is the adaptation of Occam's razor principle to science. From the set of *consistent* models $\mathcal{M}$ of observations of objects E we should choose the one which is the shortest:

$$\arg \min_{M \in \mathcal{M}} I(M) \qquad M \text{ is consistent with } E,$$

where $I(x) = -\log_2 P(x)$ denote the information of event x.

[1]This work was supported by the Ministry of Science and Technology of Republic of Slovenia (Project J2-6187), European Union Copernicus Program (Grant 1068 RECCAD), and by U.S. – Slovene Joint Board (Project #95-158).

[2]According to Bertrand Russel, the actual phrase used by William of Ockham was: "It is vain to do with more that can be done with fewer."

Fisher's *maximum likelihood principle* says that for given data E the model M which maximizes the posterior probability of data given model should be selected:

$$\arg \max_{M \in \mathcal{M}} P(E|M) = \arg \min_{M \in \mathcal{M}} I(E|M) \ .$$

We may observe several characteristics provided by empirical data or considerations based on symmetry probabilistic laws etc., which can be used as constraints in determining the model given data. Usually, these constraints are not sufficient to determine the distribution of models. E.T. Jaynes proposed the *maximum entropy principle* which is used to select the appropriate prior distribution of models given constraints. Maximum entropy selects the prior probability of models $p_i = p_{M_i}$ which maximize the entropy function and is consistent with obtained constraints

$$\arg \max H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad p_i \text{ is consistent with constraints}$$

For example, consider a loaded die ($n = 6$). If we observed that the average throw gives the average a, we have two constraints

$$\begin{aligned} \sum_{i=1}^{n} i\, p_i &= a \\ \sum_{i_1}^{n} p_i &= 1, \end{aligned}$$

which must be taken in consideration in maximizing the entropy. Observe that if we selected any prior distribution of models which would not maximize the entropy given constraints we would decrease entropy (i.e. add information) without justification.

Rissanen [12] advocates the use of information content of observations relative to a model which is called *Minimum Description Length (MDL) principle*. According to the MDL principle not only the accuracy of the model given data but also the complexity of the model should be taken in the consideration when selecting the appropriate model of the data. The MDL principle balances two factors: the encoding of a model and the encoding of data given the model. The selected model minimizes the sum of the encoding of the model and the data given model:

$$\arg \min_{M \in \mathcal{M}} \{I(M) + I(E|M)\}$$

In the process of growth (specialization) of a model the number of misclassified instances of data decreases and so does the encoding of data given refined model; on the other hand, specialization results in the increase of the model encoding, since the length of a model increases.

It can be proven (see [8], pp. 316-317) that both, the maximum likelihood principle and the maximum entropy principle are special cases of the MDL principle.

The paper describes the use of MDL principle in selecting appropriate models of

objects. In Sect. 2 MDL principle is briefly presented. It is shown that the ML (maximum likelihood) principle is a special case of MDL principle. Finally, in Sect 3 an approximate encoding for non graph theoretic models are presented. A greedy algorithm for model selection is presented in Sect 4. The algorithm takes the line segments obtained by the Hough transform [13] on an edge image and eliminates unnecessary edges based on the MDL principle.

2 MDL Principle–an Overview

Since there will generally be several models that explain objects we need a sound basis for grading models. From the set of all possible models $\mathcal{M}$ we shall choose the most probable model. Let $P(M)$ be the probability of model M in the set of all possible models (the definition of $P(M)$ will be discussed in Sect 3) and let $P(E)$ be the probability of object E in the given scene(s). If we assume that the above events are independent, we can express the probability of the model M *given* object E in the scene(s) using Bayes' theorem:

$$P(M|E) = \frac{P(E|M)\,P(M)}{P(E)}.$$

Since the probability $P(E)$ is constant for all $M \in \mathcal{M}$ it follows that the ordering of probability of models depends only of $P(E|M)\,P(M)$. If we express this product using information instead of probabilities it follows immediately that the most probable model given E is the one which minimizes the expression:

$$\arg \min_{M \in \mathcal{M}} \{I(E|M) + I(M)\} \ . \tag{1}$$

Equation 1 balances two factors: $I(E|M)$ the information needed to encode observations E given model M which decreases when M gets more specialized; but on the other hand this causes the increase of $I(M)$, the information needed to encode M itself, and vice versa. Using Bayes' Theorem we have

$$I(E|M) + I(M) = I(M|E) + I(E) \ . \tag{2}$$

Using (2), we can rewrite (1):

$$\arg \min_{M \in \mathcal{M}} \{I(E|M) + I(M)\} \quad = \quad \arg \min_{M \in \mathcal{M}} \{I(M|E) + I(E)\}$$

Since $I(E)$ is constant (the set of scenes and objects in them is fixed), we have

$$\begin{aligned} \arg \min_{M \in \mathcal{M}} \{I(E|M) + I(M)\} &= \arg \min_{M \in \mathcal{M}} \{I(M|E)\} \\ &= \arg \max_{M \in \mathcal{M}} \{P(M|E)\} \end{aligned} \tag{3}$$

which states that Rissanen's MDL principle actually maximizes $P(M|E)$ over $\mathcal{M}$.

The purpose of concept formation is *information compression;* a model describes or "explains" given data. A model M *compresses* data E if

$$I(E) > I(M) + I(E|M).$$

The most compressible hypothesis is the one with minimal encoding length and with the greatest posterior probability (see Eq 3).
Since results from algorithmic information theory have shown that finding an optimal encoding of a model is equivalent to the halting problem [9], a decidable coding scheme approximation must be adopted for calculating $I(M)$. The problem of calculating $I(M)$ will be discussed in detail in Sect. 3.

3 Approximate Encoding of Models

If we want to apply the MDL principle to model selection we have to compute, according to Eq 1 the probabilities $P(E|M)$ and $P(M)$. The former is usually easy: if we have enough scenes and objects the relative frequency of the success of recognizing object E among all objects in the scene(s):

$$P(E|M) = \frac{correct(M,E)}{correct(M,E) + incorrect(M,E)}$$

is a good approximation.
The evaluation of $P(M)$ (i.e. $I(M)$) is more difficult. How to define the probability (the information) of a model? The exact formulation is beyond the scope of this paper (see [8] for the details); let us just state that the information of the model equals the length of the shortest program that produces a model. Since finding the shortest program that produces a model is undecidable (see [9]), we have to approximate $I(M)$. Therefore we need to encode a model as well as we can to obtain a good approximation of $P(M)$.
The importance of simple explanations has a long history in modeling visual data. Gestalt psychologists summarized their observations in a number of Gestalt principles, one of them being the law of Prägnanz, or the minimum principle, which states that the visual field will be organized in the *simplest* or *the most likely* possible way [4]. Recently, simplicity in terms of the MDL principle has found its applications in computer vision [10, 7, 2, 5, 11]. The MDL principle was proposed, for example, to select the appropriate scale of observing visual data [15].
In the following subsection we encode the information needed to encode a non graph theoretical model. The program that decodes the model is omitted since it is assumed to be fixed and known in advance.

3.1 Model Encoding

In the following we restrict ourselves to modeling binary edge images (the approach can be easily extended to intensity images). Following the MDL principle,

we seek to compress the image by using models for description of the image. To evaluate the compression we propose a model for the encoding of intensity images and two models for binary images which enable us to determine the compression of image using a model. The MDL principle will also be used for searching for the most probable model of the image from the set of possible models.
An $m \times n$ intensity image with k levels of intensity can be encoded as a message of length $m \times n$ with k code symbols. Using coding based on stochastic complexity [6], the code length for such a message is:

$$I(E) = L(f_1, \ldots, f_k) = \sum_i f_i \ln \frac{f}{f_i} + \frac{k-1}{2} \ln \frac{f}{2} + \ln \frac{\pi^{k/2}}{\Gamma(k/2)} \tag{4}$$

where $f_i > 0$ is the frequency of symbol (intensity) i in the image, and $f = \sum f_i$.
For example, the encoding of $m \times n$ binary image with p zeroes takes

$$L(p, m \times n - p).$$

Let us now consider an alternative way of encoding a $m \times n$ binary image. If we apply an edge detection algorithm the resulting line segments can serve as the model of the image. Every line segment can be encoded by the coordinates of its end points. If the image is presented by k line segments, we have to encode $2\,k$ points in the $m \times n$ image which takes (see Eq 4)

$$I(M) = L(2\,k, m \times n - 2\,k). \tag{5}$$

Every black point in the image is modeled by exactly one line segment. If there are s line segments and $f_i > 0$ points is modeled by the i-th line segment, the information needed to encode the points to segments mapping takes

$$L_{p,s} = L(f_1, \ldots, f_s). \tag{6}$$

Let $P_s = \{p_1, \ldots, p_{f_i}\}$ be the set of points which belong (i.e. are modeled by) to the segment s defined by points A and B. (see Fig. 1).
Using line segment s to model the point $p_j = (x_j, y_j)$ we only need to encode x_j since y_j can be computed using the equation of the line segment s. Note that in the case that the line segment s does not accurately model the point p_j, the error $\delta_{s,j}$ must also be encoded to determine y_j. Every point in our model is represented by its x coordinate and the error of its corresponding line segment model

$$p_j = (x_j, y_j) = (x_j, k\,x_j + \delta_j + n)$$

where line segment s is defined as $y = k\,x + n$. To encode points belonging to segment s we first need to encode the x coordinate for every point. We may choose the segment $s = (A, B) = (x(A), y(A), x(B), y(B))$ in such a way that all x coordinates of points in P_s fall within $[x(A), x(B)]$. Clearly, the x coordinate of every point in P_s can be mapped to $X'_s = [0, x(B) - x(A) + 1]$. Let $X_s = \{f_{s,x} \mid x \in X'_s \wedge f_{s,x} > 0\}$ be the non-zero part of frequency distribution

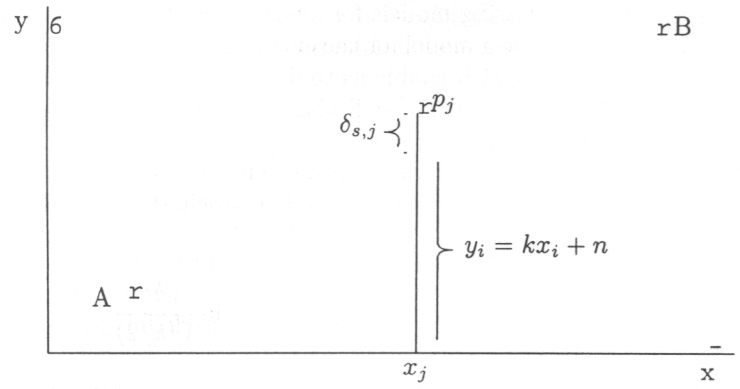

Figure 1: A set of points $p_1, \ldots, p_j, \ldots, p_{f_i}$ modeled by line segment (A, B)

of x coordinates of points in P_s mapped onto X'_s. To encode the x coordinates of points in P_s using segment s and mapping X'_s we need

$$L_{s,X_s} = L_{\{f_{s,x_i} \in X_s\}}(f_{s,x_i}) \tag{7}$$

bits of information.
The set of corrections of a line segment model s of points in P_s is coded in a similar fashion as the coding of x coordinates of points in P_s. Let Δ_s be the non-zero part of frequency distribution of corrections of model s of P_s. To encode the corrections needed to determine P_s using line segment s as a model of points in P_s we need

$$L_{s,\Delta_s} = L_{\{f_{s,\delta_i} \in \Delta_s\}}(f_{s,\delta_i}) \tag{8}$$

To summarize, the information needed to encode the points P_s given model s is the sum of encodings given in Eqs 7-8. The encoding of the complete model M of image of p points which consists of s line segments is the sum of the encodings of the constituent line segment models with addition of the mapping of points to line segments encoding

$$I(M) = L_{p,s} + \sum_{i=1}^{s} (L_{i,X_i} + L_{i,\Delta_i})$$

4 Algorithm

The implementation of the MDL principle to modeling of binary edge images with line segments is relatively simple. According to Eq 1 we search for minimal (i.e. most compressive) model of the image. In our case candidate models are sets of line segments. The input image is a typical result of processing an intensity image with an edge finder operator. Such edge images must be typically "cleaned"

before higher level processing such as stereo matching or recognition can start. Typically such "cleaning" operations consist of: thinning of edges, filling small gaps, linking of edge elements, different kinds of filtering etc. Running the Hough transform on an "uncleaned" edge image results in a multitude of overlapping lines. Different methods, for example cluster analysis [14], are proposed to select a subset of lines obtained by the Hough transform. We propose to use the MDL principle to select from this multitude of possible line models only the necessary ones. The algorithm starts with the full model (all line segments resulting from the Hough transform) then applies a local peak detection and a treshold in Hough space. Each edge point is then assigned to the closest line segment according to Euclidian distance. Line segments that do not contain enough points (usually 15) are deleted so that the remaining ones can be handled by the algorithm. We need to reduce the number of line segments because our algorithm has time complexity of $O(n^2)$, where n is the initial number of line segments. The main algorithm gradually refines the model by removing line segments from the model. It stops when there is no refined model which would be more compressive as the current model. We are not interested in global optimization over a set of possible models since this would be intractable. Besides, the experiments show that local optimization produces satisfactory results.
The algorithm performs a greedy search in the space of possible models minimizing the sum of the encoding of the model and of the image given the model. In every iteration of the algorithm we remove the line segment which results in the most compressive model. Let $M = \{l_1, \ldots, l_n\}$ be the original model of the image E consisting of n line segments. We obtain the approximation of the most compressive model as

1. Repeat
2. $C = I(M) + I(E|M)$ { complexity of the current model }
3. For every line segment $l_k \in M$
4. $M' = M - \{l_k\}$
5. If $I(M') + I(E|M') < I(M) + I(E|M)$
6. Let $M = M'$
7. Until $C > I(M) + I(E|M)$

5 Experiments

We performed several experiments on finding segment models of binary edge images. The initial model of a binary edge image was obtained by performing the Hough transform. The transform usually results in a model with many redundant line segments. When the above selection algorithm proceeds, the majority of line segments are removed. The snapshots of the process are given in Fig. 2.

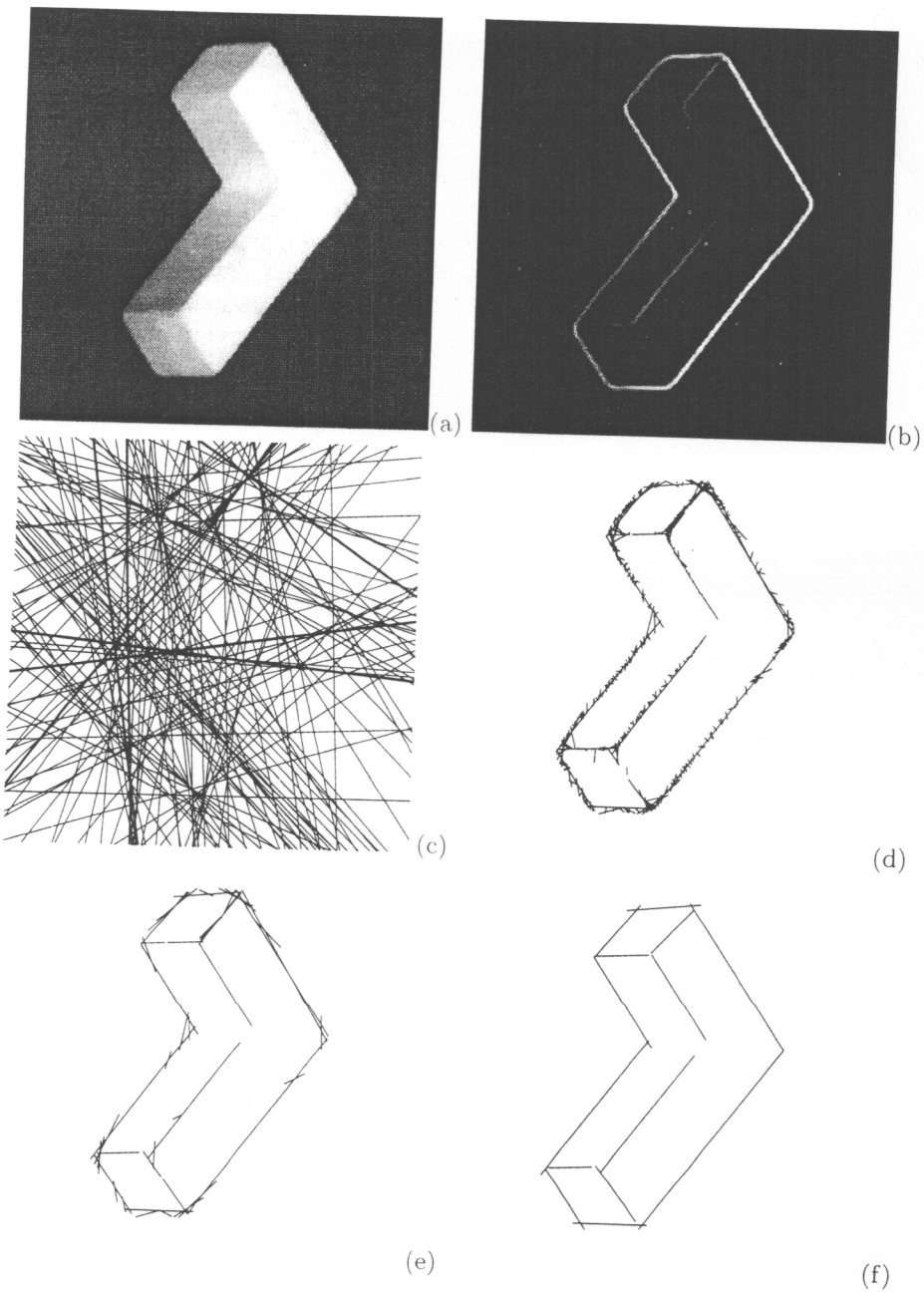

Figure 2: The algorithm performance: (a) original image (b) edges found by an edge operator (c) line segments found by the Hough transform (d) broken up line segments (e) remaining line segments after deleting those with less than 15 points (f) final model of the image

On noisy images the Hough transform produces many spurious lines. Scattered points which are far from, but on the same line as actual edges, cause the extension of the corresponding line models (see Fig. 2(c)). Therefore, we may use the above algorithm also for searching the modifications of the line segments produced by Hough transform before the actual selection of line segments.
The algorithm for breaking such line models is completely the same as described above, but instead of refining the model by eliminating line segments, we break them into more parts to obtain better models of the image. Results are shown in Fig 2(d). Fig. 2(e) shows the line segments remaining after deletion of the ones, containing less than 15 points and Fig. 2(f) the final result.

6 Conclusions

Minimum description length (MDL) principle can be used as a method for model construction from sensory data. The application of the MDL principle requires the computation of information content of the model. The paper describes the generic encoding for a non graph theoretic model. For demonstration, modeling of edges with straight line segments was performed. This example demonstrates that instead of heuristic approaches such low-level image processing tasks can be founded on sound theoretical basis.
The issue of time complexity of object reconstruction is briefly addressed. It is proposed that models with various time complexity and accuracy should be used to achieve optimal time complexity along with high reconstruction accuracy.

References

[1] R. T. Chin and C. R. Dyer. Model-based recognition in robotic vision. *ACM Computing Surveys*, 18:67–108, March 1986.

[2] P. Fua and A. J. Hanson. Objective functions for feature discrimination. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1596–1602, Detroit, MI, 1989.

[3] E. Grimson and T. Lozano-Perez. Model-based recognition and localization from sparse range and tactile data. *International Journal of Robotics*, 18:67–108, March 1986.

[4] J. Hochberg. *Perceptual Organization*, chapter Levels of Perceptual Organization, pages 255–276. Lawrence Erlbaum Associates, New Jersey, 1981.

[5] K. C. Keeler. Map representations and optimal encoding for image segmentation. Technical Report CICS-TH-292, Center for Intelligent Control Systems, March 1991.

[6] R. E. Krichevsky and V. K. Trofimov. The performance of universal coding. *IEEE Transactions on Information Theory*, IT-27:199–207, 1981.

[7] Y. G. Leclerc. Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, 3:73–102, 1989.

[8] M. Li and P. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York, 1993.

[9] S. Muggleton. *Course on Inductive Logic Programming*, 1993.

[10] A. P. Pentland. Part segmentation for object recognition. *Neural Computation*, 1:82–91, 1989.

[11] M. Pilu and R.B. Fisher. Recognition of geons by parametrically deformable contour models. In R.Cipolla and B.Buxton, editors, *Fourth European Conference on Computer Vision*, volume I of *Lecture Notes in Computer Science*, Berlin, April 1996. Springer-Verlag.

[12] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, 1984.

[13] A. Rosenfeld and A. C. Kak. *Digital Picture Processing*, volume 2. Academic Press, Orlando, FL, 1982.

[14] M. J. Silberman and J. Sklansky. Toward line detection by cluster analysis. In K. Voss, D. Chetverikov, and G. Sommer, editors, *Computer Analysis of Images and Patterns*, pages 117–122, Berlin, 1989. Akademie-Verlag.

[15] Franc Solina and Aleš Leonardis. Selective scene modeling. In *Proceedings of the 11th International Conference on Pattern Recognition*, pages A:87–90, The Hague, The Netherlands, September 1992. IAPR, IEEE Computer Society Press.