Selective Scene Modeling*

Franc Solina and Aleš Leonardis
Computer Vision Laboratory, Faculty of Electrical Engineering and Computer Science
University of Ljubljana
Tržaška 25, 61001 Ljubljana, Slovenia
franc.solina@ninurta.fer.yu, ales.leonardis@ninurta.fer.yu

Abstract

In this paper we propose an efficient architecture for selective image modeling. We give an example in which models of different scale are reconstructed in parallel. We show that this redundant representation can effectively be pruned using the criterion of Minimum Description Length. *Models that are selected in the final description indicate the appropriate scale of observation.*

1 Introduction

It is becoming clear in the vision community that a single, universal all-purposeful representation is not feasible. Such a representation would have to be excessively flexible with many parameters and hence computationally unstable. Biological visual systems and experiments are showing us that instead of one complex model, several elementary models are better encoded in a behavior. Instead of a universal representation we should try to find a robust universal vision architecture that consists of several modules, that can combine different sources of information and readily adapt itself to specific goals. These ideas are known in literature as *active* [2] and *purposive* [1] vision. Hence, a thorough (comprehensive) reconstruction of shape and depth from different visual cues in a general purpose vision system is neither necessary nor sensible, and maybe not even possible to perform. Instead, *selective reconstruction* or *modeling* should be made which would use models tailored to the task at hand.

We would like to stress in particular the relation between the goal of the vision system and the selection of appropriate models (i.e. mobile robot—even, flat surfaces, free of obstacles; grasping—volumetric models relating to the construction and size of the gripper). Reasoning about functionality in vision should be achieved through usage of specific models adapted to walking, driving, grasping, sitting etc. instead of a general purpose representation.

Reconstruction should be selective in the sense that it is made only (a) where models are applicable, (b) to a degree that is necessary for accomplishing a particular task. To represent different aspects of an image we need different models (for shape, texture, color, specularity, etc.). Besides, reconstruction must take into consideration resource and time constraints. In real situations a vision system must devote all its computing resources only to those parts of the scene that are relevant for accomplishing its goals.

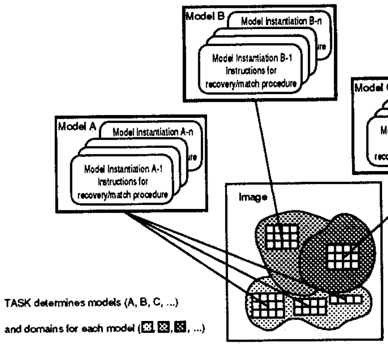

Figure 1: Architecture for selective scene modeling.

As an ultimate goal we would want to have a visual architecture as shown in Fig. 1 where each task (defined by an outside agent) determines a set of preferred models together with the appropriate image domain for each model type. The system then searches for instances of those models in the respective image domains where several instances of each type of model start detection in their corresponding seed regions. Several models of either parametric [7] or rigid [9] nature can be chosen and searched for in *parallel*. Un-

*This research was supported by the Slovenian Ministry of Research and Technology (Project P2–1122).

like "classical" segmentation that attempts to describe the whole image with a particular type of models, this scheme recovers only those models that are applicable in a specific task at hand (*domain of applicability*). Parts of the image where no model is applicable remain undescribed. Apparently, those parts of the image are not important for accomplishing the task of the system. On the other hand, some parts in the image may be, at least partially, matched by multiple models, resulting in a redundant description of the image. To get a concise description information from all recovered models, an efficient selective procedure has to be designed. We described such a procedure for selecting among overlapping models of the same kind concurrently with the recovery process, where the only difference between the models was their initial spatial position in an image, in [7,8].

In general, models can possess a variety of mutually exclusive or inclusive pieces of knowledge. In this paper we propose a novel idea on combining multiple sources of information through MDL[1] (or a similar measure) with a special emphasis on how this principle automatically determines the scale on which phenomena are (to be) observed. Models vary in the allowable deviation from the model, which results in a set of interpretations that encompass different scales[2]. An example is presented that demonstrates how both the type of the models and the design of a criterion function determine the outcome of the selection procedure.

2 Selection of Models Operating on Multiple Scales

Consider, for example, Fig. 2. How many dots in a line does one perceive as individual dots and when does the perception switch to a representation of a line (a)? How strong must the directional discontinuity of connected straight line segments be to perceive it as a whole (one line) or as individual line segments (b)? When do we consider a checkerboard pattern as a unity and when as an explicit composition of squares of different colors (c)? The importance of various representations (models) of an object has been emphasized by Bobick and Bolles [3], however the *prescribed* decision mechanism that is responsible for switching between different representations is based on an absolute criterion, namely, only when a currently used model fails, a different model is tested. In our paradigm the complexity of individual descriptions are taken into account—multiple representations are compared on a relative basis.

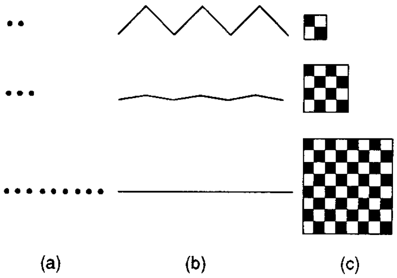

Figure 2: When does the representation switch from modeling single elements to modeling the overall structure?

The task of combining and selecting among recovered models is a problem of selecting the final interpretation among several hypothetical descriptions of the data and is solved on the level of models rather than on the level of their constituent elements (data). The goal is to minimize an objective function which has a general form:

$$F(\mathbf{m}) = \sum_i m_i \mathcal{L}_{m_i}(\mathbf{m}) \ , \qquad (1)$$

where m_i denotes a *presence variable* having the value 1 for the presence of the particular model and 0 for the absence of the model in the final description. $\mathcal{L}_{m_i}(\mathbf{m})$ denotes the weight of a particular model, which involves the complexity of the model, its spatial extent, its goodness-of-fit to the data, and the interaction of the model with the models that (partially) extend over the same domain.

The core of the problem is in defining criteria for optimality of the selected set of models. Beside dependency of models on the task we believe that there are mechanisms that operate on a more general level in order to reduce the number of redundant descriptions. Since there are computational constraints on what information can be computed in finite time and memory the available resources must be used to extract the most important information for further processing (bounded rationality—Simon [11]). Intuitively, this reduction in complexity of a representation coincides with a general notion of simplicity (Gestalt principles, i.e. law of Prägnanz—the visual field will be organized

[1] Minimum Description Length

[2] This notion of multiple scale modeling is different from the standard multiresolution which is normally based on filtering the original image with a set of different spatial operators to obtain a hierarchy of images of different resolutions. Here the original image remains intact, only the extent and allowable deviations of models are changed.

in the *simplest* or *the most likely* possible way [5]). In information science Shannon [10] revealed the importance of relation between the probability theory and the shortest encoding (simplicity). Recently, simplicity in terms of MDL principle has found its applications in computer vision [9,6,4].

Thus, our goal is to select a description that minimizes Equation (1) under the condition that portions of the image that **can** be described (at least one of the models is applicable) **must be** described (with at least one model).

An Example

Let us illustrate the MDL principle with an example. Consider a one-dimensional signal $g(t)$ which was formed by a uniform sampling of a piecewise-constant function $f(t)$ corrupted by additive independent identically distributed (IID) Gaussian noise $N(0,\sigma)$ Let N denote the number of samples. The amplitude of each sample is quantized to one of the integer values between 0 and A, and Gaussian noise is rounded to the nearest integer—the precision of the signal values. Two out of an infinite set of possible descriptions of the signal are:

1. A pointwise description. Since any integer l can be encoded with about $\log_2 l$ or $\lg l$ bits, the total length of encoding equals approximately to $N \lg A$.

2. Description in terms of a language $\mathcal{L}$. The language[3] $\mathcal{L}$ involves two components: a deterministic one which specifies the n intervals with a constant amplitude $(A_1, \ldots, A_n)$, and a stochastic one which encodes the residuals between the model and the data. The encoding necessary to describe the model requires

$$L_M = (n-1)\lg B + n \lg A \approx n \lg AB \quad \text{bits} \ , \qquad (2)$$

where the interval boundaries are quantized between 0 and B. The lowest bound on the number of bits that is required to describe data generated by a stochastic process is the negative lg of the probability of observing that data. Assuming that the residuals are modeled by an independent identically distributed (IID) Gaussian noise, $N(0,\sigma)$, the obtained optimal code will be of length

$$L_{M|D} \approx N(\lg \sigma + \frac{1}{2}\lg 2\pi e) \ . \qquad (3)$$

The total length to encode both the model and the data is thus

$$L_M + L_{M|D} \approx n \lg AB + N(\lg \sigma + \frac{1}{2}\lg 2\pi e) \ . \quad (4)$$

For this particular example, minimizing Equation 1 would answer the question which *model* (language) is better in the MDL sense?

- $\mathcal{L}_{m_1} = N \lg A$ **or**
- $\mathcal{L}_{m_2} = n \lg AB + N(\lg \sigma + \frac{1}{2}\lg 2\pi e)$.

The solution of this optimization problem will support our intuitive thinking that encoding is efficient if the number of data points described by a model is large, standard deviation low, while at the same time keeping the complexity of the model small.

Using Models of Different Scales

Let us now apply this theory to the problem of selecting models that operate on different scales. For the sake of simplicity we will assume only two types of models.

1. Model $\mathcal{M}_1$ involves two components: parameter v_1 which specifies the constant amplitude of the corresponding signal and the deviations from the value v_1. The model can describe only those intervals, which do not contain data points with a deviation exceeding Δ (let $\Delta \rightarrow 0$).

2. Model $\mathcal{M}_2$ involves two components: parameter v_2 which specifies the constant amplitude of the corresponding signal and the deviations from the value v_2. The model can describe only those intervals, which do not contain data points with a deviation exceeding δ (let $\delta \rightarrow \infty$).

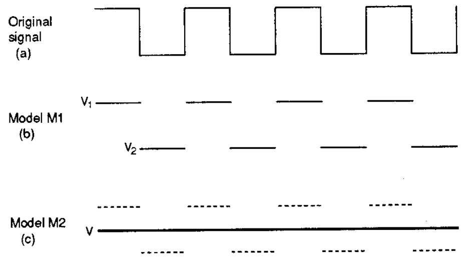

Figure 3: (a) Original 1-D signal, (b) $\mathcal{M}_1$, (c) $\mathcal{M}_2$.

For example, if we have a 1-D signal shown in Fig. 3 we can show that the above defined criteria can decide whether this 1-D signal should be represented as a piecewise composition of small segments, or as a single segment:

1. The description of the signal **in terms of the model** $\mathcal{M}_1$: The signal is partitioned into n intervals. Each of them is modeled by a constant value (V_1 or V_2). The encoding necessary to describe the signal involves only the cost of specifying the amplitude since

[3] The terms language and model mean the same in this case.

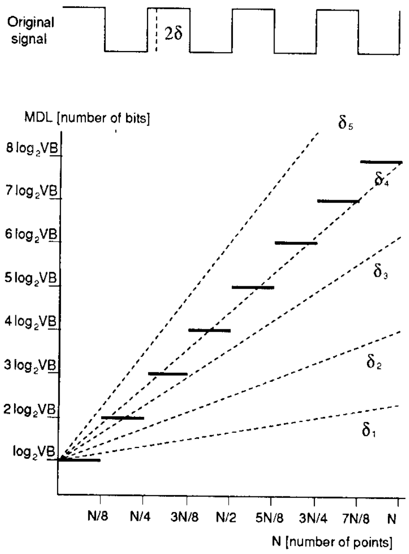

Figure 4: Relation between MDL, signal length (N), δ, and the type of the models.

there are no deviations from the model.

$$L_1 = L_M = (n-1)\lg B + n \lg V \approx n \lg VB. \quad (5)$$

B denotes the resolution of the interval boundaries.

2. The description of the signal **in terms of the model** $\mathcal{M}_2$: In this case we have to specify both the amplitude of the signal V and the deviations from its value. The total length to encode both components is

$$L_2 = L_M + L_{M|D} = \lg VB + N \lg \delta \ . \quad (6)$$

Fig. 4 shows the relation between the length (N) of the signal and the corresponding number of bits required to properly encode the signal for different deviation (δ) from the constant amplitude. The thicker lines correspond to the piecewise model. The dashed line denotes the number of bits required to model the signal as a single line plus deviations in each segment. Several dashed lines are shown for different values of δ. Interestingly enough, the results are in accordance with the intuitive expectations and can be summarized as follows:

- For no deviations or small deviations (δ_1, δ_2), the signal is represented as a single straight line, regardless of the number of points (length of the signal).
- When deviations are significant (δ_5) the signal is always described piecewise.
- In between (δ_3, δ_4), we see that the optimal selection of the model (in the MDL sense) depends on the length of the signal. For greater lengths, the signal is treated as a single model, whereas for shorter the optimal description is piecewise.

If a different encoding system (language, models) is chosen the result can change accordingly.

3 Conclusion

We propose selective image modeling on different scales in parallel. Such redundant representation can be efficiently pruned using the criterion of Minimum Description Length. Models that are selected in the final description indicate the appropriate scale of observation according to the principle of *bounded rationality.*

References

[1] Y. Aloimonos. Purposive and qualitative active vision. *Proceedings of the 10th ICPR*, pages 346–360, Atlantic City, NJ, June 1990.

[2] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8), August 1988.

[3] A. F. Bobick and R. C. Bolles. Representation space: An approach to the integration of visual information. *Proceedings of the IEEE Computer Society Conference on CVPR*, pages 492–499, June 1989.

[4] P. Fua and A. J. Hanson. Objective functions for feature discrimination. *Proc. of the 11th IJCAI*, pp. 1596–1602, Detroit, MI, 1989. Morgan Kaufman.

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

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

[7] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation as the search for the best description of the image in terms of primitives. *Proceedings 3rd ICCV*, pages 121–125, Osaka, Japan, December 1990. IEEE.

[8] A. Leonardis and R. Bajcsy. Finding Parametric Curves in an Image. *Proc. ECCV-92*, Springer-Verlag, LNCS-Series Vol. 588, 1992.

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

[10] C. Shannon. A mathematical theory of communication. *Bell Systems Tech. Journal*, 27:379–423, 1948.

[11] H. Simon. Theories of bounded rationality. In C. McGuire and R. Radner, ed., *Decision and Organization*, ch. 8, pp. 161–176, North-Holland, 1972.