AN OPTIMIZATION APPROACH FOR USING CONTEXTUAL INFORMATION IN COMPUTER VISION

Olivier D. Faugeras
Image Processing Institute
University of Southern California
Los Angeles, California 90007, U.S.A.

ABSTRACT

Local parallel processes are a very efficient way of using contextual information in a very large class of problems commonly encountered in Computer Vision. An approach to the design and analysis of such processes based on the minimization of a global criterion by local computation is presented.

INTRODUCTION

The problem of assigning names or labels to a set of units/objects is central to the fields of Pattern Recognition, Scene Analysis and Artificial Intelligence. Of course, not all possible names are possible for every unit and constraints exist that limit the number of valid assignments. These constraints may be thought of as contextual information that is brought to bear on the particular problem, or more boldly as a world model to help us decide whether any particular assignment of names to units makes sense or not.

Depending upon the type of world model that we are using, the problem can be attacked by discrete methods (search and discrete relaxation) or continuous methods (continuous relaxation). In the first case our contextual information consists of a description of consistent/compatible labels for some pairs, or more generally n-tuples of units. In the second case the description includes a numerical measure of their compatibility that may or may not be stated in a probabilistic framework. Initial estimates of likelihoods of name assignments can be obtained from measurements performed on the data to be analyzed. Usually, because of noise, these initial estimates are ambiguous and inconsistent with the world model. Continuous relaxation (also sometimes called probabilistic relaxation or stochastic labeling) is thus concerned with the design and study of algorithms that will update the original estimates in such a way that ambiguity is decreased and consistency (in terms of the world model) is increased.

More precisely, let us denote by $\mathcal{U}$ the finite set of N units and by $\mathcal{L}$ the finite set of M possible labels. In the discrete case, the world model consists of an n-ary relation $R \subset (\mathcal{U} \times \mathcal{L})^n$. The fact that the n-type $\{(u_1,\ell_1),\ldots,(u_n,\ell_n)\}$ belongs to R means that it is valid to assign name ℓ_i to unit u_i for $i=1,\ldots,n$. In the continuous case, the world model consists of a function c of $(\mathcal{U} \times \mathcal{L})^n$ into a closed interval [a,b] of the real line:

$$c: \quad (\mathcal{U} \times \mathcal{L})^n \to [a,b]$$

In most applications $[a,b]=[0,1]$ or $[-1,1]$ and $n=2$. The numbers $c(u_1,\ell_1,\ldots,u_n,\ell_n)$ measure the compatibility of assigning label ℓ_i to unit u_i for $i=1,\ldots,n$. Good compatibility is reflected by large values of c, incompatibility by small values.

We will present in this paper two ways of measuring the inadequacy of a given labeling of units with respect to a world model and show that these measures can be minimized using only local cooperative computation. We will compare this approach with the original probabilistic relaxation scheme proposed by Rosenfeld, Hummel and Zucker [3] and a matching scheme proposed by Ullman [6]. To conclude the section, we will discuss the possibility of using Decentralization and Decomposition techniques to alleviate the curse of dimensionality and show how the Optimization approach can be extended very easily to the analysis of multilevel, possibly hierarchical, systems.

We will not discuss in this paper any specific application. For an early application to scene analysis and discussion of some of the issues addressed in this paper, see [2]. For recent surveys, see [1] and [4]. For an application to graph matching, see [18].

I. Basic Optimization Based Probabilistic Relaxation Scheme

We assume that attached to every unit u_i are measures of certainty $p_i(\ell)$, for ℓ in $\mathcal{L}$ that can be thought of loosely as probabilities

$$\sum_{\ell \text{ in } \mathcal{L}} p_i(\ell) = 1 \qquad (1)$$

This work was supported in part by the Defense ARPA contract F-33615-76-C-1203. The author is with the Image Processing Institute and Department of Electrical Engineering, USC, Los Angeles, California 90007.

The world model is embedded in a function c mapping $(\mathcal{U}\times\mathcal{L})^2$ into [0,1]. Again, $c(u_1,\ell,u_2,m)$ measures the compatibility of calling unit u_1,ℓ and unit u_2,m. This function also allows us to define a topological structure on the set of units by assigning to every unit u_i and label ℓ in $\mathcal{L}$ a set $V_i(\ell)$ of related units u_j for which these exists at least one label m in $\mathcal{L}$ such that $c(u_i,\ell,u_j,m)$ is defined.

A compatibility vector $\vec{Q}_i$ is then computed for every unit u_i that measures the compatibility in each label ℓ in $\mathcal{L}$ with the current labeling at related units in $V_i(\ell)$. The simplest way of defining $Q_i(\ell)$ is [1]:

$$Q_i(\ell) = \frac{1}{|V_i(\ell)|} \sum_{u_j \text{ in } V_i(\ell)} Q_{ij}(\ell) \qquad (2)$$

where $|V_i(\ell)|$ is the number of units in the set $V_i(\ell)$ and $Q_{ij}(\ell)$ is given by:

$$Q_{ij}(\ell) = \sum_{m \text{ in } \mathcal{L}} c(u_i,\ell,u_j,m)p_j(m) \qquad (3)$$

Loosely speaking, $Q_i(\ell)$ will be large if for many units u_j in $V_i(\ell)$, the compatible labels (that is the labels m such that $c(u_i,\ell,u_j,m)$ is close to 1) have high probabilisties, and low otherwise.

In some cases the compatibility coefficients may be given a probabilistic interpretation, that is $c(u_i,\ell,u_j,m)$ is the conditional probability $p_{ij}(\ell|m)$ that unit u_i is labeled ℓ given that unit u_j is labeled m.

The next step in designing the Relaxation scheme is to specify a way of combining the two sources of information that we can use, i.e. the initial probabilities and the contextual information, to update the label probabilities. This updating should result in a less ambiguous and more compatible overall labeling in a sense that will remain vague until later on. Rosenfeld et al. [3] proposed the following iterative algorithm: for every unit u_i and every label ℓ in $\mathcal{L}_i$ set

$$p_i^{(n+1)}(\ell) = \frac{p_i^{(n)}(\ell)Q_i^{(n)}(\ell)}{\sum_{m \text{ in } \mathcal{L}} p_i^{(n)}(m)Q_i^{(n)}(m)} \qquad (4)$$

The denominator of the right hand side is simply a normalizing factor to ensure that numbers $p_i^{(n+1)}(\ell)$ still add up to one. Intuitively, the labels ℓ whose compatibility $Q_i(\ell)$ is larger than others will see their probability increase whereas the labels with smaller compatibility will see their probability decrease.

One criticism with this approach is that it does not take explicitly into account measures of the two most important characteristics of a labeling of units, namely its consistency and its ambiguity. Faugeras and Berthod [5,7,8] have proposed several such measures and turned the labeling task into a well defined optimization problem which can be solved by local computations in a network of processors.

We saw before that we can associate with every unit u_i a probability vector $\vec{p}_i$ and a compatibility vector $\vec{Q}_i$ whose components are given by equation (2). In general, the vectors $\vec{Q}_i$ are not probability vectors in that their components do not sum to 1. This can be easily changed by normalizationand we can define:

$$q_i(\ell) = \frac{Q_i(\ell)}{\sum_{m \text{ in } \mathcal{L}} Q_i(m)} \qquad (5)$$

The vectors $\vec{q}_i$ are now probability vectors and a measure of consistency for unit u_i (local measure) can be defined as the vector norm

$$C_i = ||\vec{p}_i - \vec{q}_i|| \qquad (6)$$

where $||\cdot||$ can be any norm (in practice the Euclidean norm). Similarly a local measure of ambiguity can be defined as

$$H_i = \sum_{\ell \text{ in } \mathcal{L}} p_i(\ell)(1-p_i(\ell)) = 1-||\vec{p}_i||_2^2 \qquad (7)$$

where $||\cdot||_2$ is the Euclidean norm. Combining the two measures yields a local criterion

$$J_i = \alpha C_i + \beta H_i \qquad (8)$$

where α and β weight the relative importance we attribute to ambiguity versus consistency. A global measure can then be defined over the whole set of units by averaging the local measures. Using the arithmetic average for example, we define

$$J = \sum_{\text{all units } u_i} J_i \qquad (9)$$

The labeling problem can then be stated as follows:

$(\mathcal{P})$ given an initial labeing $\{\vec{p}_i^{(0)}\}$ of the set of units $\mathcal{U}$, find the local minimum of the function J closest to the initial conditions, subject to the constraints that the vectors $\vec{p}_i$ are probability vectors. More precisely, this implies that

$$\sum_{\ell \text{ in } \mathcal{L}} p_i(\ell)=1 \text{ and } p_i(\ell)\geq 0 \text{ for all units } u_i \qquad (9a)$$

Aside from the fact that we are now confronted to a well-defined mathematical problem which can be tackled using Optimization techniques, we are also sure that some weighted measure of inconsistency and ambiguity is going to decrease.

As pointed out in [8], one minor drawback with the definition (6) of consistency is that its minimization implicitly implies the minimization of $\vec{q}_i$ and $\vec{p}_i$ and therefore the maximization of the entropy term H_i (equation (7)). Thus there is an inherent problem with the definition (8) in the sense that consistency and ambiguity tend to go in opposite directions. One very simple way of resolving that contradiction is to define a local measure of both ambiguity and consistency as

$$J_i' = -\vec{p}_i \cdot \vec{q}_i \tag{10}$$

where • denotes the vector inner product. The definition of a global criterion proceeds now as before:

$$J' = \sum_{\text{all units } u_i} J_i' \tag{11}$$

and the labeling problem can be stated as $(\mathcal{P})$, replacing J with J'. This is similar to the minimal mapping theory developed by Ullman [6] for motion correspondence. Given two image frames with elements u_i in the first one (our units) and element k (our names) in the second one, he studied the problem of selecting the most plausible correspondence between the two frames. Defining the cost $q_i(k)$ of pairing element u_i with element k and the variables $p_i(k)$ equal to 1 if u_i is paired with k and 0 otherwise, the rephrased the motion correspondence problem as a linear programming (LP) problem by defining the cost function

$$J_u' = \sum_{\text{all element } u_i} \sum_{\text{all element } k} p_i(k) q_i(k) \tag{12}$$

which is precisely equation (11). The important difference between criteria J' and J_u' is that the costs $q_i(k)$ in (12) are not functions of the variables $p_i(k)$ whereas in (11) they are. In particular, minimizing J' is not an LP problem, in general. Nevertheless, the parallel between the two approaches is interesting and confirms that what we called the compatibility coefficients $q_i(k)$ defined in Eq. (5) are also a measure of the satisfaction/profit implied in assigning name k to unit u_i.

II. Computational Requirements: Locality, Parallelism, Convergence

As described in [7,9], imagine that we attach to the set $\mathcal{U}$ of units and the sets $V_i = \bigcup_{\ell \text{ in } \mathcal{L}} V_i(\ell)$ a simple network, that is a pair <G,R> where G is a connected graph and R a set of processors, one for each node in the graph. There is a one to one correspondence between the units and the nodes of the graph on one hand, and the nodes of the graph and the processors on the other hand. This in turn implies that there is a one to one correspondence between the neighbors of the ith processor r_i, i.e., the processors that are connected by arcs of G to r_i, and units in V_i.

As shown in [7,8], the minimization of criteria J or J' can be achieved by using only local computation. More precisely, denoting by $\mathcal{J}$ (a function of all the vectors $\vec{p}_i$) either criterion J or J', we can attach to every unit u_i a local gradient vector

$$\frac{\partial \mathcal{J}}{\partial \vec{p}_i} = F_i(\vec{p}_j) \tag{13}$$

where F_i is a function of the vectors $\vec{p}_j$ of units u_j in the set V_i of neighbors previously defined. Explicit formula for the functions F_i can be found in [4,5,7,8]. The iterative scheme is then defined as

$$\vec{p}_i^{(n+1)} = \vec{p}_i^{(n)} + \rho_n P_i \left\{ \frac{\partial \mathcal{J}}{\partial \vec{p}_i} \right\} \tag{14}$$

where ρ_n is a positive stepsize and P_i a linear projection operator determined by the constraints imposed on the vector $\vec{p}_i^{(n+1)}$, [5,7], (for example that it is a probability vector). The main point is that both functions F_i and operator P_i can be computed by processor r_i by communicating only with neighboring processors (local computation) while guaranteeing that the cost function $\mathcal{J}$ will decrease globally.

It was stated before that a large amount of parallelism can be introduced in the process of minimizing criteria J and J'. This is achieved by attaching to every unit u_i a processor r_i connected only to processors r_j attached to units u_j related to u_i. The global criterion can then be minimized by having processors r_i perform simple operations mostly in parallel while a simple sequential communication process allows them to work toward the final goal in a coordinated fashion.

If nonetheless our supply of processors is limited, we may want to split our original problem into several pieces and assign sequentially our pool of processors to the different pieces. The net result has of course to be the minimization of the original global criterion and some coordination must therefore take place.

Solutions to this problem can be found in the so-called Decomposition and Decentralization techniques which have been developed to solve similar problems in Economics, Numerical Analysis, Systems Theory and Optical Control [12,13,14,15]. Decomposition techniques proceed from an algorithm standpoint: we are confronted with a problem of large dimensionality and try to substitute for it a sequence of problems of smaller dimensionalities. Decentralization techniques take a different viewpoint: we are confronted with a global problem and have at our disposal P decision centers. The question is whether it is possible to solve the global problem while letting the decision centers solve only local problems. The structure of criteria J and J' as sums of local measures allows us to develop both types of techniques [12]. The key idea is to partition the set of units. For detailed algorithms, see [16].

III. Extension to Hierarchical Systems, Conclusions

The optimization approach presented in Section I can be extended to the case where several labeling problems are present and embedded in a pyramid or cone structure with, for example, L levels.

The different levels can be the same picture at different spatial resolutions as in [17] or represent different states of abstraction. For example the lowest level could be the edge element level, then the link level [10], then the level dealing with elementary shapes like straight lines, ellipses, cubics, etc... These different levels form a multilevel system, each level having to solve a stochastic labeling problem. Let $\vec{v}_i$ be the command vector for level i, that is $\vec{v}_i$ is a N_iM_i dimensional vector, if there are N_i units and M_i possible classes, obtained in concatenating the probability vectors $\vec{p}_j$, $j=1,\ldots,N_i$. At level i we have to minimize a criterion $J_i(\vec{v}_1,\vec{v}_2,\ldots,\vec{v}_L)$. The fact that criterion J_i depends upon the command vectors at other levels accounts for the interaction between the levels.

A natural, but not always rational, way of solving this multilevel problem is to assume that every level i (i=1,...,L) considers as given the command vectors of the other levels and minimizes its own criterion. The result is a non-cooperative equilibrium [12] or Nash point $(\vec{u}_1,\ldots,\vec{u}_L)$ verifying:

$$J_i(\vec{u}_1,\ldots,\vec{u}_{i-1},\vec{u}_i,\vec{u}_{i+1},\ldots,\vec{u}_L) \le J_i(\vec{u}_1,\ldots,\vec{u}_{i-1},\vec{v}_i,\vec{u}_{i+1},\ldots,\vec{u}_L)$$

for all i and $\vec{v}_i$. This notion can certainly be criticized because by cooperating each of the L levels can, in general, improve its situation compared with the non-cooperative case. In other words, the following situation is possible: if $(\vec{u}_1,\ldots,\vec{u}_L)$ is a Nash point, there exists another set $(\vec{u}'_1,\ldots,\vec{u}'_L)$ of command vectors such that

$$J_i(\vec{u}'_1,\ldots,\vec{u}'_L) < J_i(\vec{u}_1,\ldots,\vec{u}_L) \text{ for all i.}$$

This introduces the notion of Pareto points which, intuitively, are optimal in the sense that it is impossible to find another set of L command vectors that will decrease all criteria. It is possible to show that under very general conditions [12], Pareto points can be obtained by minimizing only one criterion! In other words if $\vec{u}=(\vec{u}_1,\ldots,\vec{u}_L)$ is a Pareto point, then there exists L positive number $\lambda_1,\ldots,\lambda_L$ such that $\vec{u}$ is a minimum of criterion

$$J(\vec{v}_1,\ldots,\vec{v}_L) = \sum_{i=1}^{L} \lambda_i J_i(\vec{v}_1,\ldots,\vec{v}_L)$$

the λ_i's can therefore be interpreted as weighting factors the different levels have agreed upon.

Another interesting possibility is to assume a hierarchical structure within the L levels, level 1 being the lowest and level L the highest. We then have a cascade of optimization problems similar to what happens in the price decentralization technique mentioned in section II, where level 1 considers $\vec{v}_2,\ldots,\vec{v}_L$ as given and computes

$$\vec{u}_1 = \min_{\vec{v}_1} J_1(\vec{v}_1,\vec{v}_2,\ldots,\vec{v}_L)$$

This defines $\vec{u}_1$ as a function of $\vec{v}_2,\ldots,\vec{v}_L$. Then level 2 solves the problem of minimizing criterion $J_2(\vec{u}_1(\vec{v}_2,\ldots,\vec{v}_L),\vec{v}_2,\ldots,\vec{v}_L)$ with respect to $\vec{v}_2$, etc...

Even though the theory of hierarchical multilevel systems is still in its infancy it has been recognized for some time now [11] that it carries the possibility of solving many difficult problems in Economics, Physiology, Biology [13,14,15], Numerical Analysis and Systems Theory [12], Optimal Control. It is clear that this theory is relevant to Image Analysis.

In conclusion, we think that probabilistic relaxation techniques will play a growing role in the near future as building blocks of more and more complex vision systems. The need to quantify the behavior of these relaxation processes will become more and more pressing as the complexity of the tasks at hand rapidly increases and the global optimization framework offers a solid basis for this analysis.

REFERENCES

[1] L.S. Davis and A. Rosenfeld,"Cooperating processes for low-level vision: a survey," TR-123, Department of Computer Sciences, University of Texas, Austin, January 1980.

[2] H.G. Barrow and J.M. Tenenbaum, "MSYS: A System for Reasoning About Scenes," Tech. Note 121, AIC-SRI Int., Menlo Park, Ca., 1976.

[3] A. Rosenfeld, R.A. Hummel and S.W. Zucker, "Scene Labeling by Relaxation Operations," IEEE Trans. on Syst., Man, and Cybern., SMC-6, No. 6, pp. 420-453, June 1976.

[4] O.D. Faugeras, "An Overview of Probabilistic Relaxation Theory and Applications," Proceedings of the NATO ASI, Bonas, France, June-July 1980, D. Reidel Publishing Co.

[5] O.D. Faugeras and M. Berthod, "Scene Labeling: An Optimization Approach," Proc. of 1979 PRIP Conference, pp. 318-326.

[6] S. Ullman, The Interpretation of Visual Motion, MIT Press, 1979.

[7] O.D. Faugeras and M. Berthod,"Improving Consistency and Reducing Ambiguity in Stochastic Labeling: An Optimization Approach," to appear in the IEEE Trans. on Pattern Analysis and Machine Intelligence, 1980.

[8] M. Berthod and O.D. Faugeras, "Using Context in the global recognition of a set of objects: an optimization approach," 8th World Computer Congress (IFIP 80).

[9] S. Ullman, "Relaxation and Constrained Optimization by Local Processes," Computer Graphics and Image Processing, 10, pp. 115-125, 1979.

[10] S.W. Zucker and J.L. Mohammed, "A Hierarchical System for Line Labeling and Grouping," Proc. of the 1978 IEEE Computer Society Conference on Pattern Recognition and Image Processing, pp. 410-415, Chicago, 1978.

[11] M.D. Mesarovic, D. Macho and Y. Takahara, *Theory of Hierarchical Multilevel Systems, Academic Press, 1970.*

[12] J.L. Lions and G.I. Marchuk, *Sur Les Methodes Numeriques En Sciences Physiques Et Economiques*, Collection Methodes Mathematiques de L'Informatique, Dunod, 1976.

[13] Goldstein,"Levels and Ontogeny,"Am. Scientist 50, 1, 1962.

[14] M. Polanyi,"Life's Irreducible Structures," Science 160, 3884, 1969.

[15] D.F. Bradley, "Multilevel Systems and Biology - View of a Submolecular Biologist," in Systems Theory and Biology (M.D. Messarovic, ed.), Springer 168.

[16] O.D. Faugeras, "Decomposition and Decentralization Techniques in Relaxation Labeling," to appear in Computer Graphics and Image Processing, 1980.

[17] A.R. Hanson and E.M. Riseman, Segmentation of Natural Scenes, in A. Hanson and E. Riseman, eds., *Computer Vision Systems, Academic Press, NY, 1978, 129-163.*

[18] O.D. Faugeras and K. Price, "Semantic Description of Aerial Images Using Stochastic Labeling," submitted to the 5th ICPR.