

# GLOBAL OPTIMIZATION OF A CONSISTENT LABELING

M. Berthod

INRIA

Sophia-Antipolis 06565 Valbonne France

## ABSTRACT

We propose here a new definition of consistency, for relaxation-optimization labeling processes, which provides a natural way to deal with the no-label problem, by eliminating probability constraints. An essential characteristic is that the labeling solution of the problem is unique and may, for a large family of world models, be found very efficiently. We focus here on the general case where consistency is a polynomial function of the values of the labels.

## I INTRODUCTION

We assume in what follows that the reader is reasonably familiar with continuous relaxation processes[7]. Given a set of units  $U = \{u_i, i=1..N\}$ , and a set of labels  $L = \{l_i, i=1..M\}$ ,  $p_{ik}$  measures the value of label  $l_k$  for unit  $u_i$ , which can be interpreted, depending on the applications, as the probability for unit  $u_i$  to belong to the  $k^{th}$  class, or as the similarity between  $u_i$  and  $l_k$ , etc... Let vector  $P_i = (p_{i1}, \dots, p_{iM})$  and vector  $P = (P_1, \dots, P_N)$ :  $P$  is a continuous labeling. We assume, without loss of generality, the  $p_{ik}$ 's are non-negative.

First, an initial labeling  $P^0$  is determined by separately classifying the units. Then a parallel and iterative procedure introduces contextual relations, known as the world model, yielding an enhanced (i.e. more consistent) labeling  $P$ . We consider here a typical world model:

- $\forall i, k: V_{ik}$ , set of  $(d-1)$ -tuples of labeled units  $v_{ij}$ , neighbours of  $u_i$  for label  $l_k$ , an element  $v_{jr}$  of  $V_{ik}$  assigns label  $l_r$  to unit  $u_{j_r}$  for  $r=1, \dots, d-1$
- $\forall i, k, j: c_{ikj}$ , joint (or conditional) probability of  $u_i$  being labeled  $k$  for the local assignment of labels defined by  $v_{ij}$ .

In that case, let

$$\forall i, k: Q_{ik} = \sum_{j \in V_{ik}} \prod_r p_{j_r} \quad (1)$$

Thus the  $Q_{ik}$ 's are polynomials of degree  $d-1$  on the components of  $P$ . The corresponding vector  $Q$  is called a compatibility vector. Obviously, for a consistent labeling,  $Q_{ik}$  and  $p_{ik}$  should be large or small at the same time.

Several updating schemes have been proposed[7,1] following this line, and mostly applied with  $d=2$ , (yielding linear  $Q_{ik}$ 's). For all of them, the  $P_i$ 's are constrained to be probability vectors, i.e.:

$$\forall i, k \quad 0 \leq p_{ik} \leq 1 \quad \forall i, \sum_k p_{ik} = 1 \quad (2)$$

This normalization is vital. Due to these many inequality and equality constraints, there are many fixed points (or extrema of a criterion), and thus  $P^0$  determines the final choice of  $P$ .

The first problem, with these methods, is that the influence of  $P^0$  on  $P$  is not well defined: it may happen that  $P$  is located very far from  $P^0$ . More importantly, they do not handle satisfyingly the no-label problem[4] which arises when a probabilistic interpretation is not relevant: one would like to allow some units not to have any label at all, mainly because no labeling would be significant for them, still finding a consistent labeling for the others.

We propose here to incorporate a measure of the confidence one may have in the labeling of any unit in a definition of consistency (semi-consistency) which combines the world model and the observations, as in[1]. For a wide scope of applications, existence, uniqueness and easy computability of a semi-consistent labeling may easily be proved.

## II DEFINITION AND INTERPRETATION OF SEMI-CONSISTENCY

We define consistency as the inner product between the labeling  $P$  and a compatibility vector

$$x_{k+(M-1)i} = p_{ik}, \quad \forall i, k \quad X = (x_1, \dots, x_v)^T$$

$C$  can be conveniently expressed by:  $C(X) = A_d X^{(d-1)}$  where  $A_d$  denotes a  $d$ -way matrix (an entry of  $A$  has  $d$  indices varying from 1 to  $v$ ), and  $X^{(d)}$   $d$  occurrences of  $X$ . It can be proved then that  $C(X)$  can always be expressed by [6]

$$C(X) = X \cdot R^{d-1} X^{(d-1)} \quad (4)$$

where  $R^{d-1}$  denotes a  $(d-1)$ -way matrix, and  $\cdot$  denotes the inner product. This justifies the equivalence mentioned earlier between symmetric consistency and consistency defined by the inner product.

Actually,  $A_d$  may be applied to  $d$  different vectors  $X_1, \dots, X_d$ . Thus  $C$  may be generalized also to these  $d$  different vectors.

$$C(X_1, \dots, X_d) = A_d X_1 \dots X_d \quad (5)$$

Similarly,  $R^{d-1}$  may be generalized to  $d-1$  different vectors exactly in the same way. A nice property is that it is always possible in (4) to choose  $R^{d-1}$  such that the following property holds

$$X_{\sigma_1} \cdot R^{d-1} X_{\sigma_2}, \dots, X_{\sigma_d} = X_1 \cdot R^{d-1} X_2, \dots, X_d \quad (6)$$

for any  $X_1, \dots, X_d$ , and any permutation  $\sigma$  of the indices. The property expressed in (6) is merely a generalization of symmetry to  $d$ -way matrices.

#### A. General Conditions for a Maximum

Let  $S^n$  denote the unit sphere of order  $n$ , i.e. the set of  $X$ 's such that:  $\sum |x_i|^n = 1$ . Let  $I^k$  be the  $k$ -way unit operator ( $\forall k$ ), i.e.:  $\forall i \quad (I^k X^{(k)})_i = x_i$ .

$S_n$  may also be defined by  $X \cdot I^{n-1} X^{(n-1)} = 1$ .

Let  $S^{n*}$  be the subset of  $S^n$  of positive vectors ( $\forall i: x_i > 0$ ). Let  $S^{n0}$  be the subset of  $S^n$  of non-negative non-positive vectors ( $\forall i: x_i \geq 0 \quad \exists i: x_i = 0$ ).

Let  $S^{n+} = S^{n0} \cup S^{n*}$ .

Obviously, the global maximum on  $S^n$  is reached for  $X$  in  $S^{n+}$ . Besides, a necessary condition for  $X$  in  $S^n$  to be an extremum is that the gradient of  $C$  at  $X$  be normal to  $S^n$ . A normal to  $S^n$  at  $X$  is given by  $I^{n-1} X^{(n-1)}$ , and the gradient, using (7), is proportional to  $R^{d-1} X^{(d-1)}$ .

Thus, the condition is indeed:

$$\exists \lambda \quad \lambda I^{n-1} X^{(n-1)} = R^{d-1} X^{(d-1)} \quad (7)$$

#### B. Conditions for the positivity of non-negative extrema.

$R$ , such that the  $R_{ik}$ 's are not linear functions of the  $p_{ij}$ 's in  $V_{ik}$ , but more generally any polynomial of maximum degree  $d-1$ . An essential difference with (1) is that the initial labeling  $P_0$  is incorporated in  $R$ ; for example, using Bayesian assumptions as in [1] suggests:

$$\forall i, k: R_{ik} = P_{ik}^0 * Q_{ik} \quad (3)$$

with fixed  $V_{ik}$ 's. In graph matching (or isomorphism), the  $V_{ik}$ 's themselves are built from the observations.

Thus, let  $C(P)$  be the consistency criterion:  $C(P)$  is a polynomial of degree  $d$ , on the  $N \times M$  variables  $p_{ik}$ . This definition directly corresponds to the support function proposed in [5]. Another way to define a strictly equivalent (as shown in the next section) consistency is to start from the symmetric consistency as proposed in [1].

Classical relaxation-optimization methods start from  $P_0$  and search for a local extremum of  $C$  meeting the constraints of (2). We suggest here to relax these constraints, and to replace them by a unique normalization constraint:

$$\sum_{i,k} p_{ik}^n = 1 \quad \text{for some } n$$

That is, we look for the extremum of  $C(P)$  on the positive unit sphere of order  $n$ .

Now let  $P^*$  be the extremum, and  $\forall i \quad \lambda_i = \sum_k p_{ik}^*$ .  $\lambda_i$  may be interpreted as the confidence attached to  $P_i^*$ . If it is small, then no label is very likely for unit  $u_i$ , which means that locally consistency is low. On the contrary, if  $\lambda_i$  is large, then one, or may be several labels are highly consistent with the context, for unit  $u_i$ . Normalizing the  $P_i^*$ 's by the  $\lambda_i$ 's after convergence (and not, this is crucial, after every iteration, as it was the case in standard relaxation) yields of course probability vectors for every  $u_i$  if needed.

\* In the special case where  $C(P)$  is quadratic,  $P$  is an eigenvector of the corresponding symmetrized matrix, and conditions for the existence, uniqueness, and computability of such a positive vector have been dealt with in [2]. In the general case considered here, we exhibit similar conditions for existence and uniqueness of an extremum of  $C(P)$  on a suitable unit sphere of order  $n$ , provided  $n \geq d$ , and  $C$  is a homogeneous polynomial. Detailed proofs of the results given below are readily available, but beyond the scope of this paper.

#### III EXISTENCE AND UNICITY OF A POSITIVE SEMI-CONSISTENT LABELING

For the sake of clarity, let  $v = N \times M$ , and:

First of all, if any entry in  $A$  is strictly positive, then any entry in  $B^{d-1}$  is also strictly positive, and

$$X \in S^{n+} \Rightarrow B^{d-1} X^{(d-1)} \in S^{n*} \quad (8)$$

In most applications, this condition is (fortunately) not true. But we proved that a sufficient (and easier to meet) condition for any non-negative extremum to be positive is that:

$$X \in S^{n0} \Rightarrow Z(B^{d-1} X^{(d-1)}) > Z(X) \quad (9)$$

where  $Z(X)$  is the number of zero coordinates of  $X$ . Going back to the interpretation of labelings, the underlying idea is roughly that any labelled unit  $(u_i, k)$  indirectly influences at time any other one  $(u_j, k)$ , by string effect. It is generally easy to meet this condition.

#### C. Uniqueness of the maximum.

Let  $N^n$  denote the operator scaling any vector to the vector parallel on  $S^n$ . We proved that, provided (8) holds, and  $n \geq d$ , for any small enough  $Y$ :

$$C(N^n(X+Y)) < C(X)$$

thus, the only extrema in  $S^n$  are maxima (no minima, no saddle-points).

But, for any  $X$  in  $S^{n0}$ , we know that the gradient is in  $S^{n*}$ . Under these conditions, it can be proved that the existence of several maxima would involve at least a saddle-point or a minimum, which is not possible.

#### IV COMPUTATIONAL ASPECT

Classical optimization of a function is performed by gradient or Newton-like methods. These ones work satisfyingly here as the maximum is unique. But it turns out (though we could not prove it) that replacing at every step  $X$  by  $N^n(A X^{(d)})$  provides the fastest convergence. This is equivalent to following the gradient to infinity before projection, and is a straightforward generalization of the well-known power method for finding the maximal eigenvector in the linear case.

#### VI CONCLUSION

The method described in this paper has only been applied so far in the quadratic case, for which uniqueness is much simpler to prove. It has given satisfying results on multispectral image segmentation[3]. Experimental validation in that case comes from the following observation: when checking, near convergence, the labelings  $P_i$  which are the most stable, i.e. for which the order of the  $p_{ik}$ 's change the least, the ones with the largest confidence values  $(\lambda_i)$  are found. Moreover, there usually is for these

units one label much higher than the others.

#### References

1. M. Berthod, "Definition of a Consistent Labeling as a Global Extremum," Proc. of the 6th ICPR, pp.339-401 (1982).
2. M. Berthod, Semi-Consistency: a Solution to the no-Label Problem, proposed to CVPR, Washington (1983).
3. A. Boulanouar, Deux aspects de traitement d'images: modelisation et classification, These de docteur-ingenieur, Univ. de Paris-6, Paris, France (1981).
4. O.D. Faugeras, "Relaxation Labeling and Evidence Gathering," Proc. of the 6th ICPR (1982).
5. R.A. Hummel and S.W. Zucker, "On the Foundations of Relaxation Labeling Processes," TR80-7, Comp. Vision and Graphics Lab., Dept of Elec. Eng., Mac Gill Univ., Montreal.
6. P.M. Prenter, "On Polynomial Operators and Equations," in Non-Linear Functional Analysis and Applications, ed. Rall L.B., Academic Press, New York (1971).
7. A. Rosenfeld, R. A. Hummel, and S. W. Zucker, "Scene Labeling by Relaxation Operations," IEEE Trans. on SMC Vol. 6(6) (1976).