IMAGE SUPER-SEGMENTATION: SEGMENTATION WITH MULTIPLE LABELS FROM SHUFFLED OBSERVATIONS

Jorge S. Marques[a] *Mário A. T. Figueiredo*[b]

[a]Instituto de Sistemas e Robótica [b]Instituto de Telecomunicações
Instituto Superior Técnico, Lisboa, Portugal

ABSTRACT

This paper addresses an image labeling problem, in which it is assumed that there are multiple sensors available at each pixel with some of them possibly inactive. In addition to not being known which sensors are active or inactive, the sensor measurements are also obtained in random unknown order. Given these incomplete observations, we wish to identify which sensors are active at each site and which observations were produced by each sensor. This labeling problem extends classic image segmentation, since it allows multiple labels (*i.e.*, region overlapping). The paper provides methods to solve this problem in two scenarios: *known* and *unknown* sensor models. A new minimization algorithm, inspired by hierarchical clustering, is introduced to minimize the energy function resulting from the proposed inference criterion.

1. INTRODUCTION

Many image processing problems are formulated as inference problems in which we observe a set of noisy measurements and wish to retrieve a label (*image segmentation*) or the *true* value of the underlying image (*image restoration*) [6]. In both problems, there is a one to one correspondence between observations and image pixels.

This paper addresses a different problem, in which several observation mechanisms (sensors) are available at each pixel (multiple observations), but only a subset of them is active. We do not know which sensors are active at each pixel and the observations are shuffled, *i.e.*, we do not know which observation was generated by which sensor. Several problems can be considered in this observation scenario, such as retrieving an underlying original image from the set of multiple, incomplete, and shuffled observations. In this paper, we focus on the following problem: identifying which sensor produced each observation. This is a labeling problem, which extends classical image segmentation in the sense that each pixel is allowed to have several labels, *i.e.*, to belong to several "segments". In fact, we wish to find subsets of the image domain associated to the different sensors, but naturally these subsets will in general not be disjoint (they will overlap), since we allow pixels to have observations from multiple sensors. We will refer to this problem as *super-segmentation*.

To the best of our knowledge, the problem addressed in this paper has not been previously considered in the image processing literature.

2. PROBLEM FORMULATION

Consider a set of image sites/nodes $S = \{s_1, \ldots, s_n\}$ equipped with a neighborhood system $N : S \rightarrow 2^S$ (satisfying the standard conditions [11]: $\forall\, s \in S,\ s \notin N(s)$, and $\forall\, s, s' \in S,\ s' \in N(s) \Rightarrow s \in N(s')$. At each site, we have a set of observations, $Y_s = \{y_{s,1}, \ldots, y_{s,m_s}\}$, each of them being generated by a different mechanism (sensor). We assume that there are a total of M observation mechanisms (sensors) but only a subset of them is active at each site: $1 \leq m_s \leq M$. Furthermore, the order by which the sensors are read is unknown and may be different for different sites. If we denote by $W_s = (w_{s,1}, \ldots, w_{s,M})$ the M-tuple of outputs of the M sensors at site s, then the observed data Y_s is obtained by shuffling a subset of W_s; formally,

$$Y_s = A_s W_s \tag{1}$$

where $A_s = (a_{ij}^s)$ is a $m_s \times M$ binary matrix made of m_s randomly chosen rows of a permutation matrix P_s (doubly stochastic binary matrix), thus satisfying

$$a_{ij}^s \in \{0,1\}, \quad \sum_{j=1}^{M} a_{ij}^s = 1, \quad \text{and} \quad \sum_{i=1}^{n_s} a_{ij}^s \in \{0,1\}.$$

Although the observations appear randomly, we assume that each sensor tends to be active for several neighboring sites; *i.e.*, if a sensor, say z, is active at a site s, the probability of z also being active in the neighborhood $N(s)$ is high. This spatial dependency assumption will be formalized by modeling the labels of active sensors as a Markov random field (MRF) [6], [11].

We will consider two different problems.

1. In the first problem, we assume that the M sensors are characterized by *known* probability distributions. The observation $y_{s,i}$ is a random variable with conditional probability distribution $p(y_{s,i}|x_{s,i})$, where $x_{s,i} \in \{1, \ldots, M\}$ is the label of the active sensor that produced the i-th observation at site s. In this case, we have multiple sensors with unknown state of activity.
2. In the second problem, we considered that the sensor models are *unknown* and may be *space-dependent*; *i.e.*, it may vary along the image. To allow inference, we make the following additional assumption: that observations generated by the same sensor change slowly in space (*spatial continuity*). Therefore, if s' and s are neighboring sites and $y_{s,i}$ and $y_{s',j}$ are observations generated by the same sensor at those sites, then $y_{s,i}$ and $y_{s',j}$ have similar values with high probability; in more formal terms,

$$(s' \in N(s) \wedge x_{s,i} = x_{s',j}) \Rightarrow (y_{s,i} \approx y_{s',j}). \tag{2}$$

The estimation of the unknown labels $x_{i,s}$ can be seen as clustering problem with a spatial prior imposed by underlying neighborhood structure.

This work was supported by FCT (plurianual funding) through the PIDDAC Program funds and by project PTDC/EEACRO/098550/2008.

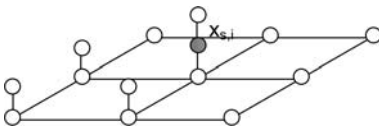

Fig. 1. Label graph: horizontal grid defines the site locations, each column represents the labels of different observations at the same location. Many neighborhood links are not represented. Each node in a column should be connected to all the nodes of the 4 neighboring columns

2.1. Known models

Let $L = \{1, ..., M\}$ denote the set of available sensors and $x_{s,i} \in L$ denote the label of the sensor which generated $y_{s,i}$, *i.e.*, the i-th observation at site s. We assume that the observations $Y = \{y_{s,i}, s \in S, i \in T_s\}$, with $T_s = \{1, \ldots, m_s\}$, are conditionally independent, given the sensor labels $X = \{x_{s,i}, s \in S, i \in T_s\}$, thus

$$p(Y|X) = \prod_{s \in S} \prod_{i=1}^{n_s} p(y_{s,i}|x_{s,i}). \tag{3}$$

Furthermore, we assume that X is a Markov random field on the following set of sites:

$$R = \bigcup_{s \in S} \{t_{s,1}, ..., t_{s,m_s}\} \tag{4}$$

Each image site s contributes with m_s sites to the set R (one per observation). Therefore, each site of R is identified by a pair (s, j) where $s \in S$ and $j \in T_s = \{1, \ldots, m_s\}$. Figure 1 illustrates this construction assuming that S is 3×3 grid.

A key question concerns the extension of the neighborhood system N defined on S to the new set of sites R; let this new neighborhood system be denoted as $Q : R \rightarrow 2^R$. Firstly, we should consider that pairs of sites sharing the same node of the original set S should be neighbors, that is,

$$\forall s \in S, \forall i, j \in T_s,\ i \neq j \Rightarrow t_{s,j} \in Q(t_{s,i}).$$

These links are necessary because labels at the same site of S must be compared to guarantee that there are no repetitions. Secondly, the nodes of R inherit the neighborhoods of the original set S, that is, $s' \in N(s) \Rightarrow t_{s',j} \in Q(t_{s,i})$, regardless of i and j.

Under the MRF hypothesis, the label configuration $X = \{x_{s,i}, s \in S, i \in T_s\}$ follows a Gibbs distribution

$$p(X) = \frac{1}{Z_l} \exp[-E_l(X)] = \frac{1}{Z_l} \exp\left[-\sum_{C \in \mathcal{C}} \phi_C(X_C)\right] \tag{5}$$

where E_l is the *labeling energy*, $\mathcal{C}$ is the set of all the cliques associated with the neighborhood system Q (*i.e.*, either singletons or sets of mutually neighboring sites [11]), ϕ_C is the potential of clique C, and X_C is the subset of X corresponding to the sites of clique C. Clearly, each subset of nodes of R corresponding to the same site of S constitute a clique, as explained in the previous paragraph.

The labeling energy should be chosen in such a way that label constraints are applied, namely,

- at any site, a sensor label cannot be repeated; formally, $i \neq j \Rightarrow x_{s,i} \neq x_{s,j}$.
- the number of label transitions should be kept as small as possible (label continuity in space).

These two desiderata are embodied by considering the following two types of clique potentials.

- The clique composed of the subset $\{t_{s,1}, \ldots, t_{s,m_s}\}$ of sites of R corresponding to the same node s of S are given a potential that forbids label repetition:

$$\phi_{\{t_{s,1},...,t_{s,m_s}\}}\ x_{s,1}, ..., x_{s,m_s} = \sum_{i=1}^{m_s} \sum_{\substack{j=1 \\ j \neq i}}^{m_s} \xi(x_{s,i} - x_{s,j}), \tag{6}$$

  where ξ is defined as $\xi(z) = \infty$, if $z = 0$, and $\xi(z) = 0$, if $z \neq 0$.

- The clique composed of a pair of sites of R, $\{t_{s,i}, t_{s',j}\}$, such that $s' \in N(s)$, is given the potential

$$\phi_{\{t_{s,i},t_{s',j}\}}\ x_{s,i}, x_{s,j} = -\beta\ \delta(x_{s,i} - x_{s',j}), \tag{7}$$

  where β is a positive parameter and δ is defined as $\delta(z) = 1$, if $z = 0$, and $\delta(z) = 0$, if $z \neq 0$.

Adding over all cliques of the two types defined in (6) and (7), we finally obtain

$$\begin{aligned} E_l(X) \quad = \quad & \sum_{s \in S} \sum_{i=1}^{m_s} \sum_{\substack{j=1 \\ j \neq i}}^{m_s} \xi(x_{s,i} - x_{s,j}) \\ & -\beta \sum_{s \in S} \sum_{s' \in N_s} \sum_{i=1}^{m_s} \sum_{j=1}^{m_{s'}} \delta(x_{s,i}, x_{s',j}). \end{aligned} \tag{8}$$

The posterior probability distribution $p(X|Y) \propto p(Y|X)p(X)$ can finally be obtained by combining (3) and (5), yielding $p(X|Y) \propto \exp(-E(X))$, where

$$E(X) = \sum_{s \in S} \sum_{i=1}^{m_s} \log p(y_{s,i}|x_{s,i}) + E_l(X). \tag{9}$$

A minimizer of $E(X)$ is a *maximum a posteriori* (MAP) estimate of X, given Y.

2.2. Unknown models

In the case where there are no *a priori* explicit models for the output mechanisms, the energy function has to be built without resorting to (3). Our approach is to keep the same labeling energy E_l, but to use a data energy that simply emphasizes data continuity. More specifically, if two neighboring observations were produced by the same sensor, it is likely that their values are similar. This idea can formalized as

$$E(X) = \sum_{s \in S} \sum_{s' \in N_s} \sum_{i=1}^{m_s} \sum_{j=1}^{m_s} \delta(x_{s,i} - x_{s',j}) \|y_{s,i} - y_{s',j}\| + E_l(X), \tag{10}$$

where E_l is as given in (8).

3. ENERGY MINIMIZATION

3.1. Known Models

The minimization of (9) is similar to an image segmentation problem with an MRF prior [6], [11], except that each site may have multiple observations and multiple labels. Energy minimization (equivalently, finding the MAP configuration) in MRF models is a very

initialize label configuration: X
initialize gain matrix $\Delta = (\delta_{ab})$: $\delta_{ab} = E(\mathcal{M}_{a,b}(X)) - E(X)$
repeat
- find $(p, q) = \arg\min_{(a,b):a\neq b} \delta_{ab}$
- update the label configuration: $X \leftarrow \mathcal{M}_{p,q}(X)$
- delete label q and renumber the remaining labels
- update Δ: remove the $q-th$ line and column and recalculate the $p-th$ line and column

until stopping criterion is satisfied.

Table 1. Hierarchical agglomerative labeling (HAL) algorithm

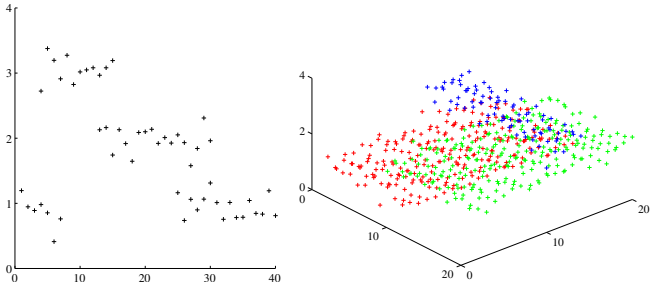

Fig. 2. 1D and 2D data obtained with multiple, incomplete and shuffle sensors

active research area in computer vision, image analysis, and machine learning, and there are many available techniques: the classical *Gibbs sampler with simulated annealing* (GS-SA) [6] and *iterative conditional modes* (ICM) [1]; the *highest confidence first* (HCF) method [3]; graph-cut (GC) based methods [7]; methods based on *loopy belief propagation* (LBP) [5], and enhancements thereof, such as the *tree-reweighted max-product* (TRW-MP) algorithm [10].

Since the goal of this paper is mostly to present a proof of concept, we consider simple and fast algorithms, namely ICM and HCF; exploring more sophisticated alternatives is left for future work. In each iteration of both ICM and HCF, a graph node (s, i) is selected; in ICM this choice follows a given (deterministic or random) schedule, while HCF has a built-in adaptive criterion to select which node to update next. The label of this node is then chosen to minimize the energy function, while keeping all other labels constant; this requires the solution of a very simple problem, since only one variable is allowed to change. However, this basic strategy is not adequate for our problem, due to the presence of forbidden label configurations at each node (recall that repeated labels are not allowed): it is not possible to move from one legal configuration to another legal configuration by changing only one label. To overcome this problem, at each iteration, rather than selecting a side $(s, i) \in R$, we select a site $s \in S$ and minimize the energy with respect to $\{x_{s,1}, \ldots, x_{s,m_s}\}$, over all legal label configurations, by exhaustive search.

Since the number of legal configurations at site s is $(M)_{m_s} = M!/(M - m_s)!$ (the Pochhammer symbol), this strategy may become too costly if M is large and m_s is not very close to 1. In future work, we will consider recent efficient algorithms which take into account forbidden label configurations [8].

3.2. Unknown Models

The minimization of (10) is a different problem, since there are no *a priori* classes available. It does not make sense to modify the labels of selected isolated sites (as in ICM or HCF). The form of (10) suggests that we should try to "connect" (assign the same label to) observations that are close in the space domain and in the feature domain. This resembles single linkage hierarchical clustering [4]; this observation underlies the algorithm that we propose to minimize (10), termed *hierarchical agglomerative labeling* (HAL). We assume that each site (s, i) is initialized with a different label, and proceed by merging pairs of neighboring nodes for which the observations are closest, and repeat this operation until a desired number of labels is reached.

Let us now formalize the proposed algorithm. Initially, each of the $\sum_{s\in S} m_s$ sites of S is given a different label. Let X be a label configuration and consider the operation of merging labels a and b (with $b > a$), meaning that every occurrence of b will be replaced by a; this operation, which will be denoted by $X' = \mathcal{M}_{a,b}(X)$, is formally defined by

$$X' = \mathcal{M}_{a,b}(X) \;\Leftrightarrow\; x'_{s,i} = \begin{array}{ll} a & \text{if } x_{s,i} = b, \\ x_{s,i} & \text{otherwise.} \end{array} \tag{11}$$

Now, let $\Delta = (\delta_{ab})$ be a *gain matrix*, where δ_{ab} represents the energy reduction associated with the merging of a and b; *i.e.*, the elements of the gain matrix are defined by

$$\delta_{ab} = E\ \mathcal{M}_{a,b}(X)\ - E(X). \tag{12}$$

After computing Δ we find the pair of labels yielding the largest energy decrease if merged (*i.e.*, the minimum off-diagonal element of Δ),

$$(p, q) = \arg \min_{(a,b):a\neq b} \delta_{ab}, \tag{13}$$

and perform the corresponding merging. Notice that for a pair of labels to be merged, two conditions have to be satisfied: **(a)** the resulting configuration is legal (no repeated labels at each site of S), otherwise $E\ \mathcal{M}_{a,b}(X)$ would be infinity; **(b)** at least a pair of neighboring sites had different labels before the merging, otherwise no energy reduction would be possible. Naturally, after each merging operation (which decreases the number of labels by 1), the set of labels should be renumbered and matrix Δ should also be rearranged to reflect this renumbering, and the procedure repeated. Fortunately, we do not have to recalculate all the elements of Δ, but only the $p-th$ line and the $p-th$ column. All the other elements of Δ are updated by adding an appropriate constant: $\delta'_{i,j} = \delta_{i,j} + C$. The algorithm is stopped if a pre-specified number of labels is reached or if the amount of energy decrease falls below some threshold. Finally, the HAL algorithm is summarized in Table 1.

4. RESULTS

The proposed model was tested with 1D data (sequences of length 40) and 2D data (20×20 grids), assuming known and unknown sensor models. The number of sensors available at each site was manually defined and the data was generated using Gaussian models, with label-dependent means $y_{s,i} \sim N(x_{s,i}, \sigma^2)$, where $\sigma = 0.2$, as shown in Fig. 2. In Fig. 2 (right), different observation mechanisms are shown with different colors to facilitate the interpretation of 3D visualization. This information (color labeling) is of course not given to the algorithm.

In the first set of experiments we assume that the sensor model is known and used the modified ICM and HCF algorithms proposed to minimize (9). The results of HCF are shown in Fig. 3. In Fig. 3 (right) we represented each of the three labels by one primary color

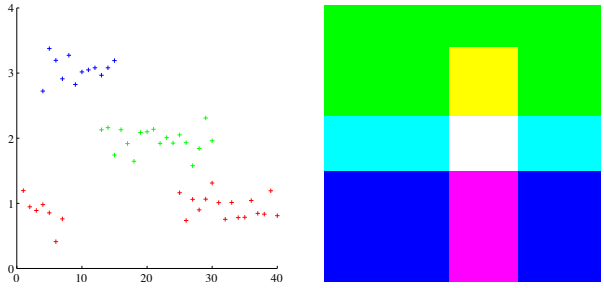

Fig. 3. Data labeling with HCF method (known sensor models)

(RGB) and multiple labels at the same site are represented by a mixture of the corresponding colors. The ICM method achieves identical results but requires a much larger number of iterations. Both methods solve the problems well, without labeling errors.

The second experiment used the same data, now assuming unknown observation models. In this case, ICM does not work well, as can be seen in Fig. 4 (1st row); the algorithm gets trapped in local minima and cannot improve the solution by changing the labels at a single site. Much better results are achieved with HCF and with the proposed hierarchical agglomerative labeling (HAL) method described in this paper. This is shown in Fig. 4 (2nd and 3rd rows). The labeling results are strongly dependent on the order in which the labels are updated. Both the HAL and HCF algorithms have their own built-in criteria to select which labels to update, and both happen to adopt a region growing strategy leading to coherent regions. On the contrary, the ICM methods adopts a random schedule, which makes it converge towards poor local minima of the energy.

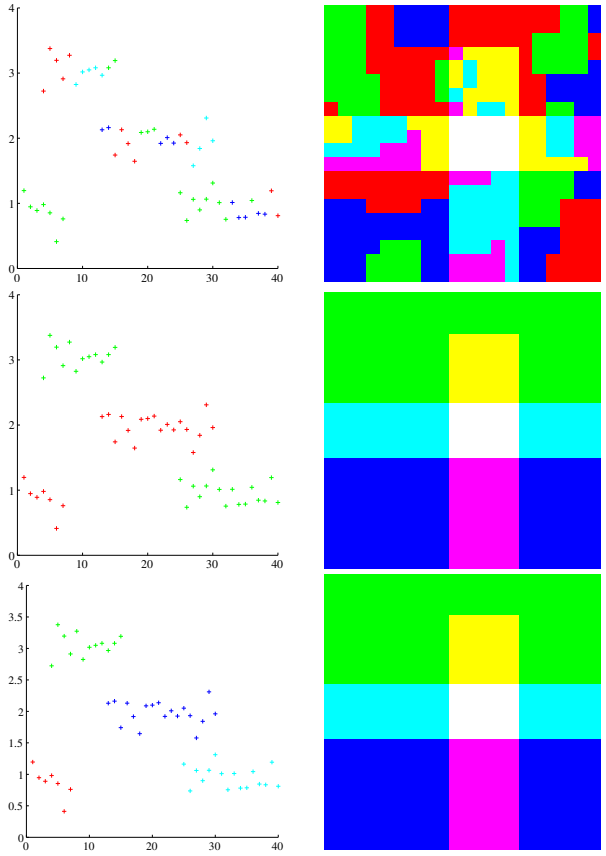

Fig. 4. Data labeling with ICM (1st line), HCF method (2nd line) and HAL method (3rd line) (unknown sensor models)

5. CONCLUSIONS

We have addressed a class of problems where we assume that each pixel is observed by multiple sensors, some of which are active and some are not. We have a (maybe space-varying) number of observations and we do not know which observations correspond to which sensors. We formulated two labeling problems for two different scenarios: **(a)** the sensor models are *known*; **(b)** the sensor models are *unknown*. These two problems were addressed in a Markov random field framework, using iterative estimation techniques. Modified versions of the ICM and HCF algorithms methods were proposed and shown to perform well in scenario (a). For scenario (b), we proposed and successfully applied a method inspired by hierarchical agglomeration of regions.

Several extensions will be addressed in future work. An obvious extension is the restoration of images assuming multiple, incomplete, and shuffled observations at each site. The application of these algorithms in the context of surveillance, namely in the estimation of multiple vector fields [9], will be addressed in another paper.

6. REFERENCES

[1] J. Besag, "On the statistical analysis of dirty pictures," *Jour. of the Royal Stat. Soc. (B)*, vol. 48, pp. 259-302, 1986.

[2] M. M. Chang, A. M. Tekalp, M. I. Sezan, "Simultaneous motion estimation and segmentation," *IEEE Trans. Image Proc.*, vol. 6, pp. 1326–1333, 1997.

[3] P. B. Chou, C. M. Brown, "The theory and practice of Bayesian image labeling," *Intern. Jour. of Computer Vision*, vol. 4, pp. 185–210, 1990.

[4] R. Duda, P. Hart, D. Stork, *Pattern Classification*, Wiley, 2001.

[5] B. Frey and D. MacKay, "A revolution: Belief propagation in graphs with cycles," *Neural Information Proc. Systems*, 1998.

[6] S. Geman, D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. and Machine Intell.*, vol. 6, pp. 721–741, 1984.

[7] V. Kolmogorov, R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Patt. Anal. and Machine Intell.*, vol. 26, pp. 147–159, 2004.

[8] A Martins, N. Smith, E Xing, P. Aguiar, and M. Figueiredo, "Augmented dual decomposition for MAP inference," *NIPS Workshop in Optim. for Machine Learning,* Whistler, 2010.

[9] J. C. Nascimento, M. A. T. Figueiredo, J. S. Marques, "Trajectory analysis in natural images using mixtures of vector fields", *IEEE Int. Conf. Image Processing*, Cairo, 2009.

[10] M. Wainwright, T. Jaakkola, A. Willsky, "MAP estimation via agreement on trees: message-passing and linear programming," *IEEE Trans. Inform. Theory,* vol. 51, pp. 3697–3717, 2005.

[11] G. Winkler, *Image analysis, random fields and Markov chain Monte Carlo methods*, Springer, 2003.