Variational Viewpoint of the Quadratic Markov Measure Field Models: Theory and Algorithms

Mariano Rivera and Oscar Dalmau

Abstract—We present a framework for image segmentation based on quadratic programing; *i.e.* by the minimization of a quadratic regularized energy linearly constrained. In particular, we present a new variational derivation of the Quadratic Makov Measure Field (QMMF) models that can be understood as a procedure for regularizing the model preferences (memberships or likelihood). We also present efficient optimization algorithms. In the QMMFs the uncertainty in the computed regularized probability measure field is controlled by penalizing the Gini's coefficient and hence it affects the convexity of the QP problem. The convex case is reduced to the solution of a positive definite linear system and, for that case, an efficient Gauss–Seidel scheme is presented. On the other hand, we present a efficient Projected Gauss-Seidel with a subspace minimization for optimizing the non–convex case. We demonstrate the proposal capabilities by experiments and numerical comparisons with interactive two-class segmentation as well as in the simultaneous estimation of segmentation and (parametric and non-parametric) generative models. We present extensions to the original formulation for including color and texture clues as well as imprecise user scibbles in an interactive framework.

Index Terms—Image segmentation, Quadratic programming, Interactive segmentation, Computer vision, Markov random fields, Information measures, Subspace minimization.

1 Introduction

Image segmentation is an active research topic in computer vision and image analysis. It is a core process in many practical applications, see for instance the listed in [1]. Image segmentation is an ill–posed problem that is task and user dependent; this is illustrated by the three possible segmentation of a single scene in Fig. 1. Among many approaches, methods based on Markov Random Field (MRF) models have become popular for designing segmentation algorithms because their flexibility for being adapted to very different circumstances as: color, connected components, motion, stereo disparity, etc.; as for example: [1], [2], [3], [4], [5], [6].

The MRFs approach allows one to express the label assignment problem into an energy function that includes spatial context information for each pixel and thus promotes smooth segmentations. The energy function codifies the compromise of assigning a label to a pixel by depending on the value of the particular pixel and the value of the surrounding pixels. Since the label space is discrete, frequently, the segmentation problem requires of the solution of a combinatorial (integer) optimization problem. In that order, max–flow/graph–cut based techniques are among the most successful optimization algorithms [7], [8], [9], [10], [11]. In particular, graph-cut based methods can solve the binary (two labels) segmentation problem in polynomial time [5]. The search for faster algorithms is, indeed, an active research topic. Recently, some authors have reported advances in the solution of the multi-label problem. Their strategy consists in constructing an approximated problem by relaxing the integer constraint [12], [13]. Additionally, there are two important issues in discrete MRFs: the reuse of solutions in the case of dynamic MRFs [8], [14] and to measure the uncertainty in the label assignement [14].

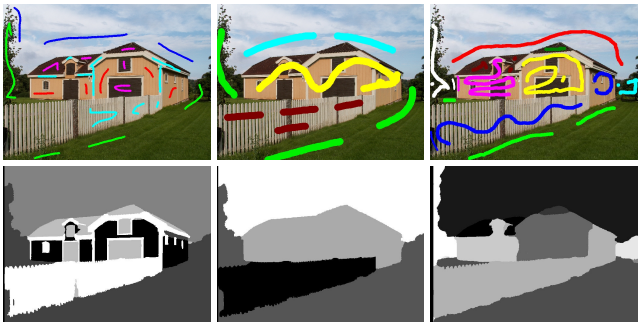

Fig. 1. Multi–class segmentation of a same scene according to different criteria (codified in the user scribbles). The columns correspond to segmentations by color, semantic objects and planar regions, respectively. The segmentation were computed with the multi-class EC–QMMF algorithm using color histograms, see section 5.1.

However, the combinatorial approach (hard segmentation) is neither the most computationally efficient nor, in some cases, the most precise strategy for solving the segmentation problem. A different approach

• *Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.*

• *The authors are with the Department of Computer Science, Centro de Investigacion en Matematicas AC , Guanajuato, GTO, Mexico 36000. E-mail: mrivera@cimat.mx, see http://www.cimat.mx/~mrivera.html*

The authors thank to the anonymous reviewers for their comments that help to improve the quality of the paper.
This work is supported by CONACYT, Mexico (Grants 61367 and 131369). O. Dalmau was supported by a CIMAT AND CONACYT scholarships.

is to directly estimate the uncertainties on the label assignment (memberships) [1], [4], [6], [15], [16]. In the Bayesian framework, such memberships can naturally be expressed in terms of probabilities—leading to the named probabilistic segmentation (PS) methods.

In this work, we present new theoretical insights, extensions and computational efficient algorithms to the recently reported PS method named Quadratic Markov Measure Field (QMMF) models [1]. In particular, we demonstrate that the data term (potential) in QMMFs is a dissimilarity measure between discrete density distributions and satisfies the here proposed design guidelines for PS methods. We also present efficient optimization algorithms proper for the two-classes (binary) and multi-classes segmentation problems. We demonstrate that the solution to a convex QMMF is computed by solving a linear system. On the other hand, since the entropy control proposed in Ref. [1] affects the convexity of the quadratic programing problem, then we propose a projection strategy combined with a subspace minimization method for the nonconvex QMMF case [17]. In addition, we include extensions to the QMMF framework that widen its capabilities.

Preliminary results of this work were reported in [18], [19], [20], [21]. We organize the paper as follows. In order to make this paper self contained, Section 2 shows a brief review of the original Bayesian derivation of the QMMF models. Section 3 presents a new variational justification of the QMMF models. The new QMMF viewpoint shows that the data term is a dissimilarity measure (an information measure) between discrete density distributions that preserves class preferences. Additionally, the presented framework demonstrates that the low entropy requirement (used in the original derivation [1]) is not a constraint in the QMMF models. Section 4 presents new efficient optimization algorithms. Then, experiments that demonstrate the method performance are presented in Section 5. Finally, our conclusions are given in Section 6.

2 Brief Review of Entropy–Controlled Quadratic Markov Measure Field Models: The Bayesian Derivation

First we introduce the notation used in this paper. We denote by r a pixel position in the image. Let $\mathcal{L} = \{r\}$ be the set of sites of a regular lattice that defines the image, then $\mathcal{R} \subseteq \mathcal{L}$ denotes the region of interest. Moreover, $\mathcal{K} = \{1, \ldots, K\}$ denotes the set of index classes and

$$\mathbb{S}^K = \{z \in \mathbb{R}^K \mid \mathbb{1}^T z = 1, z \succeq 0\} \tag{1}$$

is the simplex whose elements are probability measures; where the vector $\mathbb{1}$ has all its entries equal to one and its size is defined by the context. In our notation, given $z \in \mathbb{R}^K$ then $z \succeq 0 \iff z_k \geq 0$ for $k = 1, 2, \ldots, K$.

Recently, in Ref. [1] the Entropy Controlled Quadratic Markov Measure Field (EC-QMMF) models for image multiclass segmentation were proposed. Such models are computationally efficient and produce Probabilistic Segmentations of excellent quality. Whereas hard segmentation procedures compute a hard label for each pixel, PS approaches (as QMMFs) compute the confidence of assigning a particular label to each pixel. In the Bayesian framework, the amount of confidence (or uncertainty) is represented in terms of probabilities. In that framework $p_k(r)$ denotes the unknown probability of the pixel $r \in \mathcal{R}$ to belong to the class $k \in \mathcal{K}$. Such a vector field p is a probability measure field; *i.e.*, $p(r) \in \mathbb{S}^K$.

The QMMF formulation constructs on the generative model:

$$g(r) \quad = \quad p(r)^T I(r) + \eta(r) \tag{2}$$

where g is the observed image, the images vector $I = [I_1, I_2, \ldots, I_K]^T$ is generated with a parametric model set Φ with parameters θ. Then we use indistinctly $I_k(r)$ or $\Phi(r, \theta_k)$ for the r–pixel–value in the kth image model; where the parameters $\theta = \{\theta_k\}, \forall k$, are known or estimated. A simple example of model is $\Phi(r; \theta_k) = \theta_k + \eta(r)$. In such a case the image regions are constant planes defined by the scalar θ_k. In addition, η is a possible noise (or residual) and the probability measure $p(r)$ can be understood as a matting vector [1], [16].

In the original proposal, the QMMF models are derived from the observation model (2), assuming *i.i.d.* Gaussian noise (with zero mean and standard deviation σ) and measure vectors $p(r)$ with neglected entropy; *i.e.*, the product $p_k(r)p_l(r) \approx 0$ for $k \neq l$ at any pixel r [1]. Hence in the Bayesian regularization framework, the conditional probability of the observation given the matting factors and the image–models is given by:

$$P(g|p, \theta) \quad \propto \quad \exp\left[-\frac{1}{2\sigma^2} \sum_r \|g(r) - p(r)^T \Phi(r, \theta_k)\|^2\right] \tag{3}$$

$$\approx \quad \exp\left[\sum_r p(r)^T D_r p(r)\right] \tag{4}$$

where the approximation (4) is valid in the low–entropy limit;

$$D_r \stackrel{def}{=} \operatorname{diag}(\ -\log \hat{v}(r, \theta)) \tag{5}$$

is a diagonal matrix associated with the pixel r. Moreover, $\hat{v}_k(r, \theta) \in \mathbb{S}^K$ is the normalized probability measure vector

$$\hat{v}_k(r, \theta) \propto \exp\left[-\frac{1}{2\sigma^2} \|g(r) - \Phi(r, \theta_k)\|^2\right] \tag{6}$$

where the vector $\hat{v}$ is the normalized version of the vector v: the preferences of the data for the image–models; this leads us to the next definition:

Definition The likelihood (model preference) $v_k(r, \theta)$ is the conditional probability of observing a particular pixel value $g(r)$ by assuming that such a pixel is taken from the image I_k:

$$v_k(r, \theta) \quad \stackrel{def}{=} \quad P(g(r)|p(r) = e^k, \Phi(r, \theta_k))), \tag{7}$$

where ek is the kth canonical basis vector.

Last derivation is based on the assumption of Gaussian noise η. The generalization to other distributions, different from the Gaussian, is justified in the low entropy limit, see [1] for more details.

Following [1], given the model preferences $\hat{v}(r,\theta) \in \mathbb{S}^K$, then an effective Probabilistic Segmentation (PS) of g can be computed by solving the quadratic programing (QP) problem:

$$\min_p U(p,\theta) \quad \text{s.t.} \;\; p(r) \in \mathbb{S}^K, \quad \text{for} \;\; r \in \mathcal{R} \tag{8}$$

where the cost function has the form:

$$\begin{aligned} U(p,\theta) &= \frac{1}{2}\sum_{r\in\mathcal{R}} Q\left(p(r),\hat{v}(r,\theta)\right) \\ &\quad -\frac{\mu}{2}\sum_{r\in\mathcal{R}} \|p(r)\|_2^2 + \frac{\lambda}{2}\sum_{\langle r,s\rangle} R(p) \end{aligned} \tag{9}$$

where the scalars μ and λ are hyper–parameters that control the contribution of each term. The first term in (9) is named the data term and attaches the solution, p, to the likelihood, $\hat{v}$; the corresponding potential is given by

$$Q(p(r),\hat{v}(r,\theta)) = p(r)^T D_r p(r) \tag{10}$$

where D_r is given in (5). The second term in (9) is the Gini's index that controls the solution's entropy. The solution's entropy is penalized (promoted to be small) with $\mu > 0$ and conversely if $\mu < 0$. The third term in (9) is named the regularization one and $\lambda > 0$ promotes spatially smooth solutions. The corresponding regularization potential is given by

$$R(p) = \|p(r) - p(s)\|_2^2 \tag{11}$$

and $\langle r,s\rangle = \{(r,s) \in \mathbb{R} \mid \|r-s\| = 1\}$ denotes the set of first neighbors.

In [1], the optimum p is computed with a Projected Gauss–Seidel (PGS). Such an algorithm iterates Eqs. (16)–(18) in [1] with a clipping of negative values (projects to zero).

In addition, the QMMF models allows one the join estimation of the segmentation and the image–model parameters θ. In such a case the memberships, p, and the parameters, θ, are estimated by alternating partial minimizations until convergence:

1) $p \leftarrow \operatorname{argmin}_p U(p,\theta)$ s.t. $p \in \mathbb{S}^K$, keeping fixed θ,
2) $\theta \leftarrow \operatorname{argmin}_\theta U(p,\theta)$, keeping fixed p.

In order to guarantee convergence, it is required the descent of the global energy at each iteration, thus, for computationally efficiency purposes, these minimizations can approximately be achieved.

3 Variational Motivation for QMMFs

In the previous section is shown that, based on the Bayesian regularization framework, the QMMF energy cost (9) is derived from the observation model (2) and by assuming that the matting factor p has low entropy. Although in the low–entropy limit there exist many potentials that can approximate the conditional probability (3). For example, let $x, y \in \mathbb{S}^K$ be discrete densities, then in the low-limit entropy any of the following approximations are valid:

$$\sum_k \log(x_k^T y_k) \approx \sum_k x_k \log y_k \approx \sum_k x_k^2 \log y_k. \tag{12}$$

However, the Q–potential is preferred because it produces probabilistic segmentations of good quality and it has important algorithmic advantages. In this section, we present a study on the Q–potential that enlightens its properties and becomes unnecessary to enforce the low–entropy constraint. We also present interesting extensions to the original QMMF model.

3.1 Probabilistic Segmentation

As we have said, a Probabilistic Segmentation (PS) consists in estimating the probability measure field p such that $p_k(r)$ expresses the probability that the label kth is the correct one at the pixel r. A simple PS is given by the models preferences (likelihoods) $\hat{v}$. However, the simple addition of noise, η, in the data, g, may produce an erroneous segmentation because the probability measure field $\hat{v}$ is also noise corrupted and needs to be filtered.

If one adopts a variational approach for filtering $\hat{v}$, then a regularized energy needs to be minimized. Such an energy has, in general, two kind of terms: a data term and a regularization term. The first one attaches the regularized PS with the data (the likelihood $\hat{v}$ in this case) and the regularization term promotes a spatial smoothness. Both terms are defined by potential functions. In a classical sense, a potential that promotes "data consistency" has minimum energy when the regularized PS equals the likelihood. Hence, it is natural that those potentials are written in terms of distances (norms) or robust functions (based on M-estimators) [22]. However, we need to take into account that we deal with vectors of probability measures ($p(r), \hat{v}(r) \in \mathbb{S}^K, \forall r \in \mathcal{R}$). Thus, we can also use measures of differences between discrete densities, known as information measures. All distances are information measures but not all the information measures are distances. For example, the popular Kullback–Leibler divergence [23], [24], [25]:

$$KL(x,y) = \sum_k x_k \log \frac{x_k}{y_k}, \quad \text{with } x, y \in \mathbb{S}^K \tag{13}$$

is not symmetric, *i.e.*, $KL(x,y) \neq KL(y,x)$. Thus, we have a large set of possibilities for choosing and constructing the potentials in our regularized energy. However, we believe that the chosen information measure (potential) should fulfill a minimum requisite introduced in the following definition.

Definition *Consistence Condition Qualification* (CCQ). The potential (information measure) $M(x,y)$ preserves the CCQ if given the measure vector x, then probability

measure $x^* = \text{argmin}_x\ M(x,y)$ with $x \in \mathbb{S}^K$ satisfies: $\text{argmax}_k\ x_k^* = \text{argmax}_k\ y_k$.

If the CCQ is fullfiled for the couple of vectors x and y, for a given information measure M, then we said that x is CCQ w.r.t. y. CCQ implies that the allocation of the mode in the model preferences, y, is preserved in x^*. It means that the hard segmentation computed with a winner–takes–all (or Maximum Likelihood) estimator is undistinguished if it is acquired from y or x^*. As we said, CCQ is the minimum requirement that one should impose to the data term of a variational approach to PS. A more restricted requisite is to preserve the order preferences, see next definition.

Definition *Order Consistence Condition Qualification* (O–CCQ). The potential (information measure) $M(x,y)$ preserves the O–CCQ if given the measure vector y, then probability measure $x^* = \text{argmin}_x\ M(x,y)$ with $x \in \mathbb{S}^K$ satisfies: $x_k^* > x_k^* \iff y_k > y_k,\ \forall\, k,l \in \mathcal{K}$.

The CCQ and O–CCQ definitions are guides for designing probabilistic segmentation methods using a Variational Regularization approach; where the potentials are intuitively chosen by the algorithm designer among information measures, norms or semi-norms.

3.2 On The QMMF Data term

We note the following.

Proposition 3.1: The potential function $Q\,(x,y)$ defined in (10) is a dissimilarity measure (or information measure) between the discrete distributions x and y and preserves O–CCQ.

To prove that $Q\,(x,y)$ is an information measure we use the generalized $(\alpha,\beta,\gamma,\delta)$–information measure between two probability density functions [24]:

$$I_{(\gamma,\delta)}^{(\alpha,\beta)}(x,y) = \frac{\sum_k x_k^\alpha y_k^{\beta-\alpha} - x_k^\gamma y_k^{\gamma-\delta}}{\exp(\alpha-\beta) - \exp(\gamma-\delta)}. \tag{14}$$

with $x,y \in \mathbb{S}^K$. Then we note that (14) reduces to the Q–dissimilarity (10) when $\alpha = \gamma = \delta = 2$ and in the limit as $\beta \to 2$, a direct result of the L'Hospital's rule.

Now, we prove that the Q–dissimilarity preserves O–CCQ. First we note that D_r in (14) is a positive definite diagonal matrix. In particular, any positive definite diagonal matrix is a Stieltjes matrix (see Appendix A) and fulfill the general result stated on the next proposition.

Proposition 3.2: Let A is a Stieltjes matrix then the solution to

$$\text{argmin}_x\ \frac{1}{2}x^T A x \quad \text{s.t.} \quad \mathbb{1}^T x = 1 \tag{15}$$

is given by $x = \pi A^{-1}\mathbb{1}$; where the positive Lagrange's multiplier $\pi = (\mathbb{1}^T A^{-1}\mathbb{1})^{-1}$ acts as a normalization constant. Moreover $x \succ 0$ (is a probability measure vector). Moreover if A is a diagonal matrix then x is CCQ w.r.t. the vector composed with the diagonal of A.

The proof of proposition 3.2 is presented in the Appendix A. Then, from this proposition and noting that $(-\log y_k)^{-1} > (-\log y_l)^{-1} \iff y_l > y_k$, we can conclude that the QMMF data term preserves the order on the minimizer distributions ($x_k > x_l \iff y_k > y_l$) and hence is O–CCQ. Note that the last result is preserved for unnormalized likelihoods ($x_k > x_l \iff a y_i > a y_l$, with a scalar $a \neq 0$); *i.e.* the QMMF models can directly use unnormalized likelihoods v.

The derivation of the QMMF data term (in particular the Q–dissimilarity) presented in this section is an algebraic derivation. It is not in the sense of a Bayesian derivation where the data term is fully defined by the observation's model and the residual distribution. Indeed, both derivations are complementary, the Bayesian derivation allows us to have an initial formulation. Then such a formulation is approximated using a computational efficiency criterion. On the other hand, the algebraic derivation, based on information measures, validate the approximation used and allows us to propose new extensions.

3.3 Relationship with other information measures

For comparison purposes, we review three information measures: the Kerridge's inaccuracy, the Q-dissimilarity and the Euclidean distance. Although they are CCQ consistent there are important differences in the computed solution and algorithmic implications. In [1] is remarked that quadratic potential (10) [$\alpha = \gamma = \delta = 2$ and the limit as $\beta \to 2$ in (14)] is justified by its numerical advantage: it leads to a quadratic programming problem. However, here we show that such a selection has beneficial implications on the solution p itself.

First we analize the Kerridge's inaccuracy:

$$K(x,y) = -\sum_k x_k \log y_k. \tag{16}$$

This can be derived from (14) with $\alpha = \gamma = \delta = 1$ and $\beta \to 1$ [23], [26]. Such a potential is prone to produce hard PS with low entropy, see the next proposition.

Proposition 3.3: The solution to

$$\text{argmin}_x\ K(x,y) \quad \text{s.t.}\ \mathbb{1}^T x = 1, x \succeq 0 \tag{17}$$

is $x = e^{k^*}$; where e^{k^*} is the k^*–th vector of the standard orthonormal basis and $k^* = \text{argmax}_k\ \hat{y}_k$. Hence x is an indicator vector and holds CCQ but do not O–CCQ.

The proof is presented in the Appendix A. This result can be contrasted with the corresponding for the Q–dissimilarity: the Kerridge's inaccuracy results in a hard labeling zero–entropy solutions), this is a disadvantage due to the lack of information on the solution's confidence. In addition, the Euclidean distance

$$E(x,y) = \frac{1}{2}\|x-y\|^2 \tag{18}$$

[base of the Gaussian Markov Measure Models (GMMFs)] has the straightforward solution:

$$x = y \tag{19}$$

and evidently holds O–CCQ.

TABLE 1
Some dissimilarities between the discrete distributions p and q.

Name	Information Measure	Minimizer p given q	O-CCQ	Optimization Problem	Gaussian model parameters
Kerridge	$-\sum_x x_k \log y_k$	$x_k = \begin{cases} 1 & y_k \geq y_l, k \neq l, \\ 0 & \text{otherwise.} \end{cases}$	No	Combinatorial optimization	Easily computable
Q-dissimilarity	$-\frac{1}{2}\sum_k x_k^2 \log y_k$	$x_k = \frac{(\log y_k)^{-1}}{\sum_l (\log y_l)^{-1}}$	Yes	Quadratic Programming	Easily computable
Euclidean	$\frac{1}{2}\sum_k (x_k - y_k)^2$	$x_k = y_k$	Yes	Quadratic minimization	No appropriate

Table 1 presents a summary of the discussed information measures. We can see that both the Q-dissimililarity and the Euclidean distance lead to Quadratic optimization problems. However, the Q–dissimilarity is preferred over the Euclidean distance because it has experimentally demonstrated that produces results with lower entropy [1]. This is an important property in the case of the joint estimation of segmentation and distribution parameters, see Section 3.4.3. In contrast, the use of the Euclidean distance results in a collapse to a single model [1], [20].

3.4 Generalizations to QMMFs

3.4.1 Inter–Pixel Affinity

In this section, we introduce the inter–pixel affinity w_{rs} as a likelihood that the pixels r and s belong to the same class. Let be the quadratic regularization potential

$$R(p(r), p(s)) = \|p(r) - p(s)\|_2^2 w_{rs} \tag{20}$$

then the purpose of w is to lead the class border to coincide with large image gradients. For example, w can be computed with :

$$w_{rs} = \frac{\gamma}{\gamma + \|T\{g(r)\} - T\{g(s)\}\|_2^2} \tag{21}$$

where γ is a positive parameter that controls the edge sensibility and T is in general a nonlinear transformation that depends on the task. Usually T is a transformation of the space color for the pixel value (*e.g.* RGB space to the Lab space [18]). However the Lab-space distance (as the color human perceptual distance) hardly represents the inter–class (objects) distances. Inter-class distances are context and task dependent. For instance, if the task is to segment the image in Fig. 1 into semantic regions then the weights should be close to one in the whole house facade, independently if there are large color gradients within. Here, we propose a new inter–pixel affinity measure based on the marginal likelihoods and thus incorporates, implicitly, the non-euclidean distances of the feature space. We chose $T\{g(r)\} = \hat{v}(r, \theta)$, then w_{rs} is a prior that the pixels r and s belong to the same class.

3.4.2 Color/Texture Based Interactive Segmentation

We propose an interactive method for image segmentation with color and texture features. The purpose is to demonstrate that the final segmentation is improved by combining multiples sources (likelihood vectors). This combination of sources is naturally implemented in our proposal. The method constructs on the computation of significance degrees of color/texture features and it is based on our previous work in [27]. Such significances are used for weighting the original features. We introduce the method using color/texture descriptors the coefficients based on the discrete cosine transform (DCT) of an image patch centered at r. Such image patches have size equal to $W \times W$ with three layers (the rgb–channels). Then the feature vectors for the hand–labeled data are denoted by $\tilde{g}(r) \in \mathbb{R}^{3\times W^2}$. The method here presented is general enough and accepts others color or texture features.

Let $\mathcal{J} = \{1, 2, \ldots, W^2\}$ be the DCT coefficient index set and $\hat{v}_k^{(j)}(r)$ the normalized likelihood of pixel r to belong to the class k using *only* the $j \in \mathcal{J}$ feature (in *rgb*). Then, we assume that the confidence factor of a given feature j is its capability for predicting the correct pixels class. Such a confidence α_j is large if the model preferences $v_k^{(j)}(r)$ of the hand-labeled pixels are large for their respective models (and small for the other ones). In particular the confidence of the jth source for the kth class can be estimated with

$$\alpha_{jk} = \frac{1}{|\mathcal{R}_k|} \sum_{r \in \mathcal{R}_k} \hat{v}_k^{(j)}(r, \theta). \tag{22}$$

If the likelihoods are normalized ($\sum_k v_k^{(j)}(r) = 1, \forall j, r$) then $\alpha_{jk} = 1$ represents a high confidence on the source jth for predicting the kth label. Then the confidence of the feature j on all the classes is $\alpha_j = \sum_k \alpha_{jk}$, and its normalization:

$$\hat{\alpha}_j = \frac{\alpha_j}{\sum_l \alpha_l}. \tag{23}$$

Finally we propose to use

$$D_r = \sum_j \alpha_j D_r^{(j)} \tag{24}$$

in the QMMF data term; where D_r is the matrix in the quadratic norm in (10) and $D_r^{(j)} = \text{diag}(\ -\log \hat{v}^{(j)}(r, \theta))$ is the contribution to the energy of the jth feature at the rth pixel.

3.4.3 On the Image–Model Parameter Estimation

In Refs. [1], [20] we studied the particular case estimating the mean of Gaussian Likelihood functions. In that case

the model preferences (likelihoods) are Gaussians of the form

$$v_k(r,\theta) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left[-\frac{1}{2\sigma_k^2}(g(r)-m_k)^2\right] \tag{25}$$

then the parameter estimation step is computed with the formulas:

$$m_k = \frac{\sum_r p_k^2(r) g(r)}{\sum_x p_k^2(r)} \tag{26}$$

$$\sigma_k^2 = \frac{\sum_r p_k^2(r)(g(r)-m_k)^2}{\sum_r p_k^2(r)}. \tag{27}$$

The proof is straightforward of solving $\nabla_\theta \sum_r Q\left(p(r), \hat{v}(r,\theta)\right) = 0$ for the parameters. Excepting the precise definition of the weight, $p_k^2(r)$, formulas (26) and (27) are similar to those used in the Expectation-Maximization (EM) procedure. The class mean, m_k, computed with (26) can be understood as the mean of the data contributions to each class, k. Such contributions correspond to $p_k^2(r)$.

Inspired in last analysis we propose a procedure for updating non–parametric likelihood density functions (as Histograms or Kernel–Histograms) in the QMMF framework. The procedure is motivated in the multi-class interactive segmentation context. We assume that some pixels in the region of interest, $\mathcal{R}$, are labeled such that $\mathcal{R}_k \subset \mathcal{R}$ is the subset of labeled pixels with label k. Then we define the multimap image $\mathcal{A}$ as $\mathcal{A}(r) = k$ if $r \in \mathcal{R}_k$ and $\mathcal{A}(r) = 0$ if $r \in \mathcal{R} \backslash \bigcup_k \mathcal{R}_k$. Hence $\mathcal{A}(r) = 0$ indicates the unlabeled pixels.

Thus, let g be an image such that $g(r) \in t$, with $t = \{t_1, t_2, \ldots, t_T\}$ are the pixel values (maybe vectorial values as in the case of color images). Then the regular histogram for the kth class is computed with:

$$H_k(t) = \frac{1}{|\mathcal{R}_k|} \sum_{r \in \mathcal{R}_k} \delta(|t - g(r)|) \tag{28}$$

$$= \frac{\sum_{r\in\mathcal{R}} \delta(\mathcal{A}(r)-k)\delta(|t-g(r)|)}{\sum_{r\in\mathcal{R}} \delta(\mathcal{A}(r)-k)} \tag{29}$$

where $|\mathcal{R}_k|$ is the cardinality of the set $\mathcal{R}_k$ and δ is the Kronecker's delta function. By analogy between (26) and (29), we can understand the term $\delta(\mathcal{A}(r)-k)$ as the contribution of the r–pixel to the histogram H_k. Then in a QMMF procedure, and if an estimate of p is available, we can reestimate (update) the histograms with:

$$H_k(t) = \frac{\sum_{r\in\mathcal{R}} p_k^2(r)\delta(|t-g(r)|)}{\sum_{r\in\mathcal{R}} p_k^2(r)}. \tag{30}$$

Experiments that demonstrate this procedure are presented in subsection 5.4.

4 Minimization Algorithms

In this section we present two new efficient minimization algorithms for solving the QMMF's optimization problem. The algorithms here presented are simpler, require less memory and have faster convergence than the original QMMF algorithm reported in [1]. First, we present a discussion on the convergence of the Projected Gauss-Seidel with Sub–Space Minimization algorithm. Then we presents a memory efficient algorithm that update entry–wise the measure field p. Although the scheme is initially developed for the QMMFs convex case, we show that it can be adapted to the non-convex case using a subspace minimization strategy. Finally, we present a faster algorithm with a vector-wise update scheme.

4.1 On the Convergence of the Projected Gauss Seidel with Subspace Minimization

Consider the general QP problem

$$\min_x F(x) = \frac{1}{2}x^T A x \quad \text{s. t. } x \succeq 0 \tag{31}$$

where the Hessian matrix A is a no–definite matrix with non–zero diagonal elements. Then we iterate of two steps: i) a Projected Gauss-Seidel (PGS) update scheme and ii) a Subspace Minimization (SSM).

i) **PGS**. First, we consider the iterative update formula

$$x^{t+1} \leftarrow x^t + h u^t, \tag{32}$$

(for $t = 0, 1, 2, \ldots$, until convergence) where $u^t = -e^k \nabla F(x)^T e^k$, with $k = 1 + (t \mod K)$, is a descent direction, e^k the kth basis vector of the canonical base. The step size h is computed using the Newton formula: $h = \beta / A_{kk}$, where the constant β is chosen such that x^{t+1} is kept feasible: $x^{t+1} \succeq 0$.
Now, by construction, u^t guarantees that x_k is the only updated element, then this algorithm can be written simply as:

$$x^{t+1} \leftarrow \max\left\{0, \frac{A_{kk}x^t - e^k \nabla F(x^t)^T e^k}{A_{kk}}\right\}, \tag{33}$$

with $A_{kk} \neq 0$. This update formula is, indeed, the PGS scheme for (31).

ii) **SSM**. Let x be the current point computed the PGS, then $\mathcal{A} = \{k \in \mathcal{K} \mid x_k = 0\}$ is the current active constraint set. Thus, in order to accelerate the convergence one can combine the upgrade of x entry by entry with another strategy that takes a subset of PGS updated entries. Let $\tilde{x}$ be the elements x_i such that $i \notin \mathcal{A}$. Thus, let $\tilde{A}$ be the submatrix of A whose columns correspond to elements in $\tilde{x}$. An approximated solution to this subproblem can be found with a Gauss Seidel scheme too. However, it is more effective to use a Newton's method with step size that guarantees a feasible point.

The described PGS-SSM is, indeed an active set methods for quadratic potentials [17].

Remark. The iteration of the PGS step by itself converges to a solution (local minima). According with our

experiments, the SSM step improves significantly the convergence time.

Note that if the problem (31) includes an equality constraint of the form $Ax = b$ where the matrix A has linear independent rows, one can use a variable elimination technique, see Ref. [17], in order to write the original problem in form of (31). In our particular case, the constraint $\mathbb{1}^T p(r) = 1$ (for all r and for a selected class i) can be eliminated by substituting $p_i(r) = 1 - \sum_{k \neq i} p_k(r)$ into the energy $U(p)$.

4.2 Memory Limited Gauss Seidel Scheme

Let d be a vector field defined as

$$d_k(r) \stackrel{def}{=} -\log \hat{v}_k(r) - \mu. \tag{34}$$

Then, if the entropy control μ is chosen such that the energy (9) is kept convex ($d_k(r) > 0, \ \forall k, r$) then the computation of p consists in solving a linear system. This is stated in next proposition.

Proposition 4.1: (Convex QMMF) Let $U(p)$ be the energy function defined in (9) and assume $d_r \succeq 0$, then the solution to

$$\min_p U(p) \quad \text{s.t.} \ \ \mathbb{1}^T p(r) = 1, \quad \text{for} \ \ r \in \Omega$$

is a probability measure field: it holds $p_r \succeq 0$.

Proof: We present an algorithmic proof to this Proposition. The optimal solution satisfies the Karush-Kuhn-Tucker (KKT) conditions:

$$p_k(r) d_k(r) + \lambda \sum_{s \in \mathcal{N}_r} (p_k(r) - p_k(s))\, w_{rs} = \pi(r) \tag{35}$$

$$\mathbb{1}^T p(r) = 1 \tag{36}$$

where π is the vector of Lagrange's multipliers. Note that the KKT conditions are a symmetric and positive definite linear system that can be solved with very efficient algorithms such as Conjugate Gradient or Multigrid Gauss–Seidel. In particular, a simple Gauss-Seidel (GS) scheme results from integrating (35) w.r.t. k (i.e. by summing over k) and using (36):

$$\pi(r) = \frac{1}{K} d(r)^T p(r). \tag{37}$$

Thus, from (35):

$$p_k(r) = b_k(r)\, [\bar{p}_k(r) + \pi(r)] \tag{38}$$

where we defined:

$$\bar{p}_k(r) \stackrel{def}{=} \lambda \sum_{s \in \mathcal{N}_r} w_{rs} p_k(s) \tag{39}$$

and

$$b_k(r) \stackrel{def}{=} \frac{1}{d_k(r) + \lambda \sum_{s \in \mathcal{N}_r} w_{rs}}. \tag{40}$$

Eqs. (37) and (38) define a two step iterative algorithm. Moreover, if (37) is substituted into (38), we can note that if an initial p is chosen positive, then the GS scheme (38) will produce a convergent nonnegative sequence. □

In addition, the entropy of the solution p can be controlled by means of the μ parameter that penalizes the Gini's (entropy) coefficient. A positive μ reduces the entropy but may result in a negative value of $d_k(r)$, see (34), and hence it leads us to a nonconvex QP problem. In this case, we can use the projection strategy for enforcing the non-negativity constraint. Then at each iteration, the projected p can be computed with

$$p_k(r) = \max\{0, b_k(r)[\bar{p}(r) + \pi(r)]\}. \tag{41}$$

Algorithm 1 summarizes the PGS procedure for updating a single vector $p(r)$. The complete process repeats the PGS step for all the pixel positions until convergence.

One can see that the GS scheme, here proposed [Eqs. (37) and (38)], is simpler than the originally reported in [1].

Algorithm 1 Simple Projected Gauss Seidel for QMMF. Update procedure for the $p(r)$ vector.

1: {Require}
   i. Let K be the number of classes, $\lambda \geq 0$ the regularization parameter, $\hat{v}$ the normalized likelihood and w the intra-pixel affinity;
   ii. Given $d_k(r)$ computed with (34) ;
   iii. Let r be the current pixel position and $p \succeq 0$;
2: **for** all k **do**
3: Update $\pi(r)$ with (37);
4: Update $p_k(r)$ with (41);
5: **end for**

In order to accelerate the algorithm convergence, we combine the PGS and the Subspace Minimization strategies. First we update $p(r)$ with the PGS scheme by neglecting the nonnegative constraints. Next, at each pixel, we estimate the active set $\mathcal{A}_r$ from the non-positive coefficients in $p(r)$. Then, we refine the previous solution by fixing $p_i(r) = 0$ for $i \in \mathcal{A}_r$ and solving (8) for the remaining $p_i(r)$ with $i \notin \mathcal{A}_r$. If an updated coefficient results negative, then the active set $\mathcal{A}_r$ is updated and a new partial solution is computed. The partial solution after few subspace minimizations (we used 2 recursions in our experiments) is used as starting point for a new PGS iteration. The procedure details are in the Algorithm 2. Note that the subspace minimization (line 6) can be computed with the same algorithm, in a recursive procedure.

In Addition, the GS scheme for the binary (two classes) segmentation can be simplified with the elimination of the variable p_2 (using $p_2 = 1 - p_1$). In such a case, the GS update formula is given by

$$p_1(r) = \frac{d_2(r) + \lambda \sum_{s \in \mathcal{N}_r} w_{rs} p_1(s)}{d_1(r) + d_2(r) + \lambda \sum_{s \in \mathcal{N}_r} w_{rs}}. \tag{42}$$

Finally, we can also use the projection strategy in the non-convex case. In such a case, the projection needs to take into account both $p_1(r)$ and $[1 - p_1(r)]$; *i.e.*,

$$p_1(r) \leftarrow \max\{0, \min\{p_1(r), 1\}\}. \tag{43}$$

Algorithm 2 Non–Convex QMMF with subspace minimization

1: {Initialization}
   i. Let K be the number of classes, $\lambda \geq 0$ the regularization parameter, $\hat{v}$ the normalized likelihoods and w the intra-pixel affinity;
   ii. Compute $d_k(r)$ with (34) ;
   iii. Compute $b_k(r)$ with (40) ;
   v. Initialize $p \succ 0$; {*e.g.* $p = \hat{v}$}
2: **repeat**
3: **for** all r **do**
4: Update $p(r)$ with Algorithm 1;
5: Compute an estimate of the active set for $p(r)$: $\mathcal{A} = \{k \in \mathcal{K} \mid \tilde{p}_k \leq 0\}$;
6: Solve approximatelly (8) for $\tilde{p}_r$ fixing $\tilde{p}_{ir} = 0$ for $i \in \mathcal{A}$ and $p(s)$ for $s \neq r$.
7: Set $p(r) = \tilde{p}_r$.
8: **end for**
9: **until** convergence

4.3 Vector-wise Gauss Seidel Scheme

Since the iterative update formula (38) [and its projected version(41)] requires of a reduced amount of memory, its is proper for processing large data, as video or tomographic images (MRI or TC volumes). On the other hand, we can improve the computational performance (convergence rate) with an extra memory cost if, instead of updating $p(r)$ component by component, we update the entire vector in a single step. First we write the KKT conditions (35) for the full vector $p(r)$:

$$D_r p(r) + \lambda \sum_{s \in \mathcal{N}_r} (p(r) - p(s))\, w_{rs} \quad = \quad \pi(r)\mathbb{1}. \quad (44)$$

Note that the KKT conditions (44) and (36) still are a symmetric and positive definite linear system. Following a similar algebraic procedure as the one used in section 4.2 we have the positive definite and diagonal dominant system:

$$H_r p(r) = \bar{p}_r \quad (45)$$

with $H_r = D_r + \Lambda_r - \hat{\mathbb{1}} d_r^T$, with $\hat{\mathbb{1}} = \mathbb{1}/K$; where we define the diagonal matrix $\Lambda_r \stackrel{def}{=} \left(\lambda \sum_{y \in \mathcal{N}_r} w_{rs}\right) I$ and the kth component of the vector $\bar{p}(r)$ is computed with (39). Then, the inverse matrix H_r^{-1} can efficiently be computed with the Sherman-Morrison formula. Thus

$$H_r^{-1} \quad = \quad B_r + \frac{B_r \hat{\mathbb{1}} d(r)^T B_r}{1 - \hat{\mathbb{1}}^T B_r d(r)} \quad (46)$$

where we define the positive diagonal matrix $B_r = \mathrm{diag}[\, b(r)]$ with the elements of $b(r)$ computed with (40). Hence, we note that

$$\hat{\mathbb{1}}^T B_r d(r) = \frac{1}{K} \sum_{k \in \mathcal{K}} \frac{d_k(r)}{b_k(r)} < 1$$

because $\lambda > 0$ and then $\frac{d_k(r)}{b_k(r)} = \frac{d_k(r)}{d_k(r) + \lambda \sum_s p_k(s)} < 1$. Thus $H^{-1} \succeq 0$ (has non negative elements) is positive definite (it does not have rows equal to zero). Consequently, the iteration of

$$p(r) = H_r^{-1} \bar{p}(r), \forall r \quad (47)$$

keeps $p \succ 0$ if the initial p is chosen positive.

Now, we give an extra step for simplify the update formula. First, we note that $B_r \hat{\mathbb{1}} = b(r)/K$. Second, we define pre–computable vector (independent of p):

$$\tilde{d}(r)^T = \frac{1}{K - b(r)^T d(r)} d(r)^T B_r. \quad (48)$$

Next, we define the product:

$$\pi(r) = \tilde{d}(r)^T \bar{p}(r). \quad (49)$$

Finally, the kth component of $p(r)$ in (47) can be computed the simple formula:

$$p_k(r) = b_k(r) \left[\bar{p}_k(r) + \pi(r)\right]. \quad (50)$$

Note the similarity with (39), the difference is the formula for computing π; compare (37) with (49). The complete procedure is summarized in Algorithm 3.

Algorithm 3 Vector-wise GS for Convex QMMF

1: {Initialize as Algorithm 2.}
2: Compute $\tilde{d}(r)^T$ with (48);
3: **repeat**
4: **for** all r **do**
5: Compute $\pi(r)$ with (49);
6: Update $p_k(r)$ with (50) for $k = 1, 2, \ldots, K$;
7: **end for**
8: **until** convergence

Note that, for the non-convex case, the substitution of the PGS step in Algorithm 2 by the vector-wise update scheme may produce a sub-optimal solution. This is because the clipping of negative values in the updated vector [with (50)] does not necessary guarantee to reduce the energy. However, according with our experiments, we have found that such a suboptimal strategy is faster and produces suboptimal solutions of good quality.

5 Experiments

We have presented theoretical aspects and practical implication of the QMMF models. In following experiments we focus on demonstrating the method capabilities.

5.1 Multiclass interactive segmentation

The image segmentation task is obviously an ill-posed problem, i.e. there exist multiple"valid" segmentations for a particular image. User interaction is a popular way for introducing prior (high level) knowledge for segmenting images with complex scenes. In this paradigm the user labels by hand a subset of pixels. Then the unknown labels are estimated with a segmentation algorithm that takes into account the distribution of the

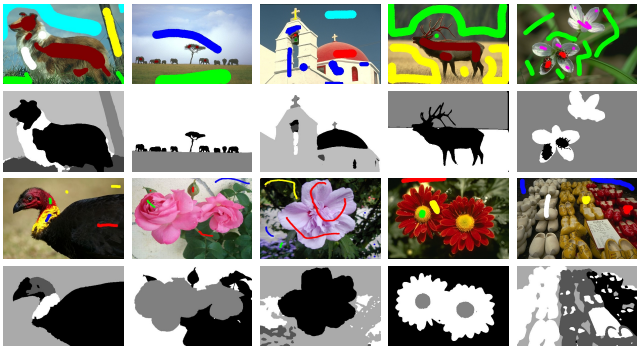

Fig. 2. Interactive multi–class segmentations.

labelled pixels and the smoothness of the spatial segmentation. Interactive segmentation is an approach that allows us to develop general purpose tools. To illustrate this, we can see three possible segmentations of the image in Fig. 1. The first column shows scribbles given by the user and the second column the corresponding segmentations computed with the method here presented. The rows correspond to segmentation by color, semantical objects (house, vegetation, fence, etc.) and planar regions; respectively.

The likelihood functions are estimated using histograms, as is described in section 3.4.3. Such histograms are computed using (29). Thus the normalized histograms are computed with $\hat{h}_{ki} = H_k(t_i)/\sum_l H_k(t_i)$ and the likelihood of the pixel r to belong to a given class k (likelihood function, LF) is computed with:

$$LF_{ki} = \frac{\hat{h}_{ki} + \epsilon}{\sum_j (\hat{h}_{ji} + \epsilon)}, \ \forall k; \tag{51}$$

with $\epsilon = 1 \times 10^{-8}$, a small constant. The ϵ scalar introduces an uniform distribution that avoids a possible division by zero and guarantee positive likelihoods. Thus the likelihood of an observed pixel value is computed with $v_k(r) = LF_{ki}$ such that $i = \arg\min_j \|g(r) - t_j\|^2$.

In this experiments, we assume that the user's labels are correct; then, in the data term in (9), the sum over all the pixels in the region of interest, $\{r \in \mathcal{R}\}$, is replaced by the sum over the unlabeled pixels; *i.e.*, for $\{r \in \mathcal{R} : \mathcal{A}(r) = 0\}$. Alternatively, the sum for all pixels supposes may be incorrectly hand labelled pixels.

Figure 2 shows multi–class interactive segmentations computed with the proposed algorithm implemented in Matlab (in *.m* and *.mex* files). Moreover, we use $T\{g(r)\} = \hat{v}(r)$ in (21).

5.2 Color/Texture based Interactive Segmentation

Texture is evidently an important clue to be consider in image segmentations. As color Likelihood Functions (LFs), feature–texture LFs can be learned from user's scribbles and can be represented with non-parametric (or parametric) models. In the case of color, it seems natural that the joint LFs are represented by 3D histograms (corresponding to the 3D color representation).

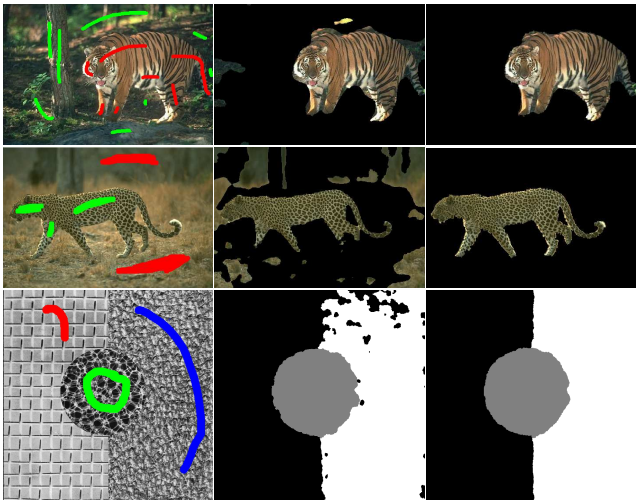

Fig. 3. Color and Texture based interactive segmentations: user's scribbles (first column), color based segmentation(second column) and color/texture based segmentation (third column)

TABLE 2
Cross-validation results: Parameters, Akaike information criterion, training and testing error.

Algorithm	Params.	AIC	Training	Testing
Graph Cut	λ, γ	8.58	6.82%	6.93%
Rand. Walk.	λ, γ	6.50	5.46%	5.50%
GMMF	λ, γ	6.49	5.46%	5.49%
QMMF	λ, γ	6.04	5.02%	5.15%
QMMF+EC	λ, γ, μ	3.58	3.13%	3.13%

Since texture features vector are in general represented in dimensions higher than 3, one can use a dimension reduction technique (as PCA) in order to find a representation with linearly independent coordinates and then use low–dimensional histograms. Different to the PCA approach, in section 3.4.2 we proposed method that finds the most significant features for segmenting a particular image: it takes advantage of the information codified in user's scribbles. In the experiment, we use the five most significant features (largest α) and 3D histograms as Likelihood Function for each cosine transform *rgb*–coefficient. According to our experiments, the use of the whole feature vector produces similar results, but its is more efficient in memory usage to use just the five more significant DCT coefficients. The Fig. 3 presents comparative results of a color–based segmentation and the proposed color/texture based segmentation.

5.3 Quantitative Comparisson: Image Binary Interactive Segmentation

Next, we summarize our results of a quantitative study on the performance of the segmentation algorithms: the proposed Binary variant of QMMF, the max–flow (min–graph–cut), GMMF and Random Walker (RW). The reader can find more details about this study in

TABLE 3
Adjusted parameters for the results in table 2.

	GC	RW	GMMF	QMMF	EC–QMMF
$\lambda(\times 10^3)$	0.128	2700	2100	4.7	230
$\gamma(\times 10^6)$	27	3.7	3.7	9.14	5700
$\mu(\times 10^3)$	—	—	—	—	−570

Fig. 4. Segmentation example from the Lasso's data set.

our technical report [19]. We used our implementation for the GMMF and RW algorithms, and the authors' implementation of the max–flow/graph–cut algorithm described in [28] for minimizing the energy:

$$U(p) = \sum_{r:A(r)=0} \Big\{ b(r)\hat{v}_1(r) + [1-b(r)]\hat{v}_2(r) + \frac{\lambda}{2} \sum_{s\in\mathcal{N}_r} w_{rs}[1-\delta(b(r)-b(s)] \Big\} \quad (52)$$

with $b(r) \in \{0,1\}$.

The task is to segment color images into background and foreground allowing interactive data labeling. The generalization capabilities of the methods are compared with a *cross-validation* procedure [25]. The comparison was conducted on the Lasso benchmark database [7]; a set of 50 images available online [29]. Such a database contains a natural image set with their corresponding trimaps and the ground truth segmentations. Actually, a Lasso trimap is an image of class labels: no–process mask ($\mathcal{M} = \mathcal{L}\backslash\mathcal{R}$), background, foreground and unknown; where no error is assumed in the initial labelled pixels. First column in Fig. 4 shows an image from the Lasso database and second column the corresponding trimap; the gray scale corresponds with the above class enumeration. In this case, the region to process is labeled as "unknown" and the boundary conditions are imposed by the foreground and background labeled regions.

We opted to compute the weights using the standard formula (21) (*i.e.* $T\{g(r)\} = g(r)$ in (21)). In order to focus our comparison on the data term of the different algorithms: QMMF, GC, GMMF and RW. In this task, empirical likelihoods are computed from the histogram of the hand labeled pixels [8]. Figure 5 shows examples of segmented objects from images in the Lasso database. One set corresponds to the results of the GraphCut (GC) considered the estate of the art and the second group to the EC-QMMF.

The hyper parameters (λ, μ, γ) were trained by minimizing the mean of the segmentation error in the image set by using the Nelder and Mead simplex descent [30]. We implement a cross–validation procedure following the recommendation in Ref. [25] and split the data set

(a) GraphCut

(b) EC-QMMF

Fig. 5. Segmented objects with GC (upper group) and EC–QMMF (lower group) algorithms.

into 5 set s, 10 images per set. Figure 4 shows an example of the segmented images. Table 2 shows summarizes the Akaike Information Criterion (AIC) [25] and the training and testing error. The AIC was computed for the optimized (trained) parameters with the 50 image in the database. Note that the AIC is consistent with the cross-validation results: the order of the method performance is preserved. Moreover the QMMF algorithm has the best performance in the group. Table 3 shows the learnt parameters for the evaluated methods. In the case of the EC-QMMF, the computational time was about 0.2 sec. for the Lasso images. This automatic learning parameter process confirms that GMMF and RW, as close variants, have similar performance [19]. However, it produces two unexpected results:

i. Our GC based segmentation improves significantly the reported results in [7]. Indeed, our basic GC formulation of the method in [8] overcomes, significantly, the reported results with Likelihood Functions based on Gaussian Mixtures [7].
ii. The learnt parameter μ (EC-QMMF) promotes large entropy. In our opinion, there are three reasons for such result: the Lasso dataset have a narrow band of unknown pixels; the trimaps are correct, so that the class models (histograms) are reliable; and the hand–segmentations (ground truth) favor smooth boundaries.

This results remark the importance of the presented variational derivation for the QMMFs that does not constraint the entropy to be small (Section 3). However,

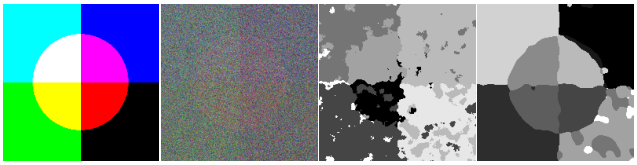

Fig. 6. Simultaneous segmentation and parameter estimation. From left to right: original image (200×200 pixels), noisy image ($\sigma = 1.8$), segmentation with $\mu = 0$ and with $\mu = 20$.

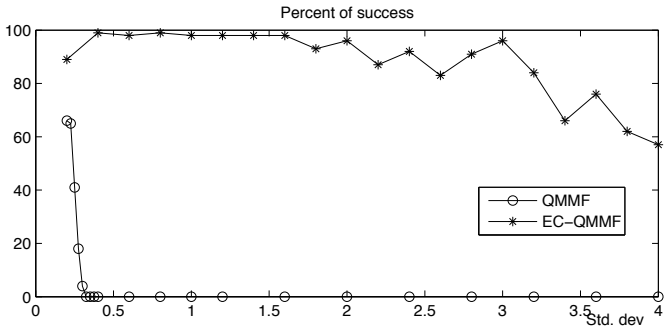

Fig. 7. Simultaneous segmentation and parameter estimation. Results on 100 Montecarlo experiments: number of correct successful detected parameters versus the SNR, see text.

next experiment demonstrates that for the simultaneous estimation of the segmentation and the model parameters the uncertainty in the solution (entropy) need to be controlled and small.

5.4 Model parameter estimation

The entropy control allows us to adapt the algorithm for different tasks. For example, lower entropy produces better results for the task of simultaneously estimation of segmentation and model parameters, see Section 3.4.3. Next, we describe an experiment that demonstrate last claim. The test image is the one shown in Fig. 6. The image has 8 regions and the colors correspond to the vertices of the rgb color–space cube. Then, such colors are the model parameter m_k, for $k = 1, 2, \ldots, 8$. The colors were normalized; *i.e.* $m_1 = [0, 0, 0]$ (black) and $m_8 = [1, 1, 1]$ (white). Then the image was corrupted with Gaussian noise with different standard deviation (SNR) levels $\{\sigma_n\}$. For each noise level, 100 Montecarlo experiments were performed. Each experiment consisted of three stages: a) to generate a random noisy image (with Gaussian noise), b) to initialize 12 means (1.5 the number of true model number) and c) to estimate the original image colors (parameters) and the segmentation with QMMF algorithm. The last step was implemented with eight iteration of the Segmentation/Parameter-estimation steps. We marked an experiment as "success" if at least a mean (of the 12) converge to each region $\mathcal{M}_k = \{z \in \mathbb{R}^K \mid \|z - m_k\|_1 < 1\}$, $\forall k$. Figure 6 shows an instance of the random images (with $\sigma = 1.8$); its

Fig. 8. Iterative estimation of empirical likelihood functions by histograms of p^2–weighted data. Binary segmentation: initial scribbles, first iteration and second iteration; respective columns.

failed segmentation with the QMMF with $\mu = 0$ (no-entropy control) and its successful segmentation with EC-QMMF, $\mu = 20$. The over–segmentation (in the EC-QMMF segmentation) can be reduced by grouping regions, but such a step is beyond the scope of this work. Figure 7 summarize the Montecarlo experiment results. The λ parameter was tuned with the thumb rule: $\lambda = 10(1 + \text{floor}(\sigma))$; then $\mu = 0$ for the non-entropy control case (QMMF) and $\mu = \lambda$ for EC-QMMF. As the QMMF ($\mu = 0$), our implementations based on Gauss Markov Measure Fields (GMMF, an early variant of Random Walker [6]) collapsed to a single model [4]. This limitation of the GMMF model is discussed in [31], see also [19].

Finally, we demonstrate the generalization of the QMMFs for the joint task of segmentation and non-parametric Likelihood Functions (LF) estimation, based on histogram techniques. The segmentation is computed after two iteration of the stages Segmentation/Parameter-estimation. First at all, the initial histograms are computed according (30) using $p = \hat{v}$. Then, the QMMF–segmentation is computed. After, the histograms are updated using (30). Finally, the QMMF–segmentation is recomputed. Different from Section 5.1, in this case the region of interest is the complete image; *i.e.*, the labels of the user marked pixels (multimap) are also estimated. The process is illustrated in Fig. 8: the erroneous segmentation after the first iteration is product of inaccurate scribbles and thus an inaccurate initial LF. The segmentation, after two iterations, demonstrate the ability of the QMMFs for estimating nonparametric class distributions.

6 Conclusions and Discussion

Image segmentation consists in partitioning an image into regions with similar features of interest (color, texture, motion, depth, etc.) or semantical properties (kind

tissue in medical images, objects in a scene, roads in aerial images, etc.). The image regions provides a compact representation that allows one to make inference on the image properties. Therefore image segmentation is an active research topic in computer vision and image analysis. It is a core process in many practical applications. In this work, we studied theoretical properties, proposed new optimization algorithms and presented practical extensions to a recent image segmentation model.

We presented a derivation of the QMMF model independent of the minimal entropy constraint. Therefore, based on prior knowledge, we can control the amount of entropy increment, or decrement, in the computed probability measures. We demonstrated that the QMMF models are general and flexible to be used with diverse Likelihood Functions. As demonstration of such a generalization, we presented experiments with iterative estimation of likelihood functions based on histogram techniques. We proposed robust likelihoods that improve the method performance for segmenting textured regions.

Our contributions in this work are mainly a more general derivation of the QMMF models and more efficient optimization algorithms. Along the paper we present a series of experiments for demonstrating our proposals. Additionally, we present an experimental comparison with respect algorithms of the state of the art. We selected the task of binary interactive segmentation for conducting our comparison, first because it demonstrates the use of the entropy control in the case of generic likelihood functions. Second, a benchmark database is online available, and finally our hyper–parameter training scheme demonstrates to be objective by, significantly, improving the previously reported results with a graph cut based method.

APPENDIX A

Definition (Stieltjes matrices [32]). A $K \times K$ Stieltjes matrix $A = (a_{ij})$ with $i, j = 1, 2, \ldots, K$ satisfies: it is symmetric and positive definite; it has positive diagonal elements, $a_{ii} > 0$; it has nonpositive off-diagonal elements, $a_{ij} \leq 0,\ i \neq j$; and its inverse $B = (b_{ij})$ is nonnegative, $b_{ij} > 0, \forall i, j = 1, 2, \ldots, K$.

Proof of Proposition 3.2: The KKT conditions of (15) are

$$Ax - \pi \mathbb{1} = 0 \tag{53}$$

$$\mathbb{1}^T x - 1 = 0, \tag{54}$$

where π is the Lagrange's multiplier. Then from (53): $x = \pi A^{-1}\mathbb{1}$. Substituting this result in (54), we have $\pi \mathbb{1}^T A^{-1} \mathbb{1} = 1$, thus $\pi = 1/(\mathbb{1}^T A^{-1} \mathbb{1})$. We conclude that $x \succeq 0$, since A is Stieltjes matrix, A^{-1} is positive definite and non-negative. Thus, $\mathbb{1}^T A^{-1} \mathbb{1} > 0$ (sum over all the elements) and $A^{-1}\mathbb{1} \succ 0$ (sum over rows). Note that A^{-1} has not rows equal zero. □

Proof of Proposition 3.3: The KKT conditions of (17) are: $d - \pi \mathbb{1} = \xi, \mathbb{1}^T x - 1 = 0, x^T \xi = 0$, and $x, \xi \succeq 0$; where $d_k \stackrel{def}{=} -\log y_k$, π is the Lagrange's multiplier of the equality constraint and ξ is the vector of Lagrange's multiplier of the nonnegativity constraints. Then, by the Fundamental Theorem of the Linear programing [17] it can be seen that the KKTs are fullfil with: $x = e_{k^*}$ with $k^* = \underset{k}{\operatorname{argmin}}\, d_k$, $\pi = d_{k^*}$ and ξ given by the first KKT. □

REFERENCES

[1] M. Rivera, O. Ocegueda, and J. L. Marroquin, "Entropy-controlled quadratic Markov measure field models for efficient image segmentation," *IEEE Trans. Image Processing*, vol. 8, no. 12, pp. 3047–3057, Dec. 2007.

[2] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. Springer-Verlag, Tokyo, 2001.

[3] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, no. 6, pp. 721–741, 1984.

[4] J. L. Marroquin, F. Velazco, M. Rivera, and M. Nakamura, "Probabilistic solution of ill–posed problems in computational vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 337–348, 2001.

[5] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts," in *European Conference on Computer Vision (ECCV02).*, 2002.

[6] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, 2006.

[7] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive GMMRF model," in *ECCV*, vol. 1, ser. LNCS 3021, Springer Berlin / Heidelberg, 2004, pp. 414–427.

[8] Y. Boykov and M.-P. Jolly, "Interactive graph cut for optimal boundary & region segmentation of objects in N–D images," in *ICIP (1)*, 2001, pp. 105–112.

[9] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[10] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Probabilistic fusion of stereo with color and contrast for bi-layer segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1480–1492, 2006.

[11] C. Rother, V. Kolmogorov, and A. Blake, "Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics*, no. 23 (3), 2004, pp. 309–314.

[12] C. Olsson, A. P. Eriksson, and F. Kahl, "Improved spectral relaxation methods for binary quadratic optimization problems," *Computer Vision and Image Understanding*, vol. 112, pp. 30–38, 2008.

[13] N. Komodakis, G. Tziritas, and N. Paragios, "Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal–dual strategies," *Computer Vision and Image Understanding*, vol. 112, pp. 14–29, 2008.

[14] P. Kohli and P. H. S. Torr, "Measuring uncertainty in graph cut solutions," *Computer Vision and Image Understanding*, vol. 112, pp. 30–38, 2008.

[15] M. Rivera, O. Ocegueda, and J. L. Marroquin, "Entropy controlled Gauss-Markov random measure fields for early vision," in *in Proc. VLSM*, vol. 1 LNCS 3752, Springer Berlin / Heidelberg, 2005, pp. 137–148.

[16] A. Levin, A. Rav-Acha, and D. Lischinski, "Spectral matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1–14, 2008.

[17] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer Series in Operation Research, 2000.

[18] M. Rivera and P. P. Mayorga, "Quadratic markovian probability fields for image binary segmentation," in *in Proc. ICCV, Workshop ICV 07*, 2007, pp. 1–8.

[19] ——, "Comparative study on quadratic Markovian probability fields for image binary segmentation," CIMAT A.C., Mexico, Tech. Rep. 10.12.2007, I-07-15 (CC), December 2007.

[20] M. Rivera, O. Dalmau, and J. Tago, "Image segmentation by convex quadratic programming," in *Int. Conf. on Pattern Recognition (ICPR08)*, 2008.

[21] O. Dalmau and M. Rivera, "A general bayesian markov random field model for probabilistic image segmentation," in *in Procc. Int. Workshop Combinatorial Image Analysis (IWCIA)*, ser. LNCS 5852, Springer Berlin / Heidelberg, 2009, pp. 149–161.
[22] M. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *Int'l J. Computer Vision*, vol. 19, no. 1, pp. 57–92, 1996.
[23] B. D. Sharma and R. Autar, "Relative information functions and their type (alpha, beta) generalizations," *Metrika*, vol. 21, no. 1, pp. 41–50, 1974.
[24] I. Taneja and H. C. Gupta, "On generalized measures of relative information and inaccuracy," *Aplikace Matematiky*, vol. 23, pp. 317–333, 1978.
[25] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, 2001.
[26] D. F. Kerridge, "Inaccuracy and inference," *J. Royal Stat. Soc, Ser. B*, vol. 23, no. 1, pp. 184–194, 1961.
[27] M. Rivera, O. Dalmau, and W. Mio, "Spatial sampling for images segmentation," in *to appear in International Symposium on Computer and Information Sciences (ISCIS)*, 2010.
[28] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Int'l J. Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.
[29] http://research.microsoft.com/vision/cambridge/i3l/segmentation/GrabCut.htm.
[30] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, pp. 308–313, 1965.
[31] J. L. Marroquin, B. C. Vemuri, S. Botello, F. Calderon, and A. Fernandez-Bouzas, "An accurate and efficient Bayesian method for automatic segmentation of brain MRI," *IEEE Trans. Medical Imaging*, vol. 21, pp. 934–945, 2002.
[32] R. S. Varga, *Matrix Iterative Analysis*, 2nd ed. Springer Series in Computational Mathematics, 2000, vol. 27.

REFERENCES

[1] H. Kopka and P.W. Daly, *A Guide to LaTeX*, third ed. Harlow, U.K.: Addison-Wesley, 1999.

Mariano Rivera received the B.E. degree in electronics from the Durango Institute of Technology, Mexico, in 1989, the M.Sc. degree in electronics from the Chihuahua Institute of Technology,Mexico, in 1993, and the D.Sc. degree in optics from the Center for Research in Optics (CIO), Leon, Mexico, in 1997. Since 1997, he has been with the Computer Science Department, Center for Research in Mathematics (CIMAT), Guanajuato, Mexico. His current interests include computer vision, image processing, machine learning, and optimization. His research is summarized in more than 70 papers in scientific journals and conference proceedings. Dr. Rivera is Fellow of the National Researcher System (SNI) of the Mexican Government.

Oscar Dalmau received his BScEd in Mathematics at ISP Manzanillo Cuba in 1989, his M.Sc. degree in Computer Science and Industrial Mathematics at the Center of Research in Mathematics (CIMAT) Mexico in 2004 and his Ph.D. degree in Computer Science at CIMAT Mexico in 2010. He works at the Center for Research in Mathematics (CIMAT), Monterrey, Nuevo Leon, Mexico. His research interests lie in the areas of machine learning, optimization, and image processing and computer vision.