Iterative Local-Global Energy Minimization for Automatic Extraction of Objects of Interest

Gang Hua, *Student Member*, *IEEE*,
Zicheng Liu, *Senior Member*, *IEEE*,
Zhengyou Zhang, *Fellow*, *IEEE*, and
Ying Wu, *Member*, *IEEE*

Abstract—We propose a novel global-local variational energy to automatically extract objects of interest from images. Previous formulations only incorporate local region potentials, which are sensitive to incorrectly classified pixels during iteration. We introduce a global likelihood potential to achieve better estimation of the foreground and background models and, thus, better extraction results. Extensive experiments demonstrate its efficacy.

Index Terms—Variational energy, level set, semisupervised learning.

1 Introduction

AUTOMATIC extraction of objects of interest (OOI) is very important in early vision with applications to object recognition, image painting, visual content analysis, etc. Given an arbitrary image, the OOI is usually subjective, but should be at the focus of attention. When one takes a picture of an OOI, one normally tries to put it roughly at the center. With this weak assumption, we are able to build a fully automatic extraction system. Note that this assumption does not tell us where the OOI boundary is.

To extract the OOI, we need to model both the OOI and background regions, e.g., the models can be Gaussian [1], Gaussian mixture (GMM) [2], [3], [4], or kernel densities [5]. The OOI extraction and the estimation of the models is a *chicken-and-egg* problem, i.e., knowing one leads to the other. When both are unknown, the problem can be solved iteratively as an energy minimization problem [1], [5], [4], i.e., fixing the models, performing the segmentation, fixing the segmentation, and re-estimating the models.

Previous variational energy formulations only incorporate local region potentials. One challenge is that at each iteration the model estimations are based on inaccurately labeled image pixels since the current segmentation is usually not perfect. The incorrect labels affect the accuracy of the estimated models which, in turn, affects the subsequent segmentation. To relieve this problem, our energy formulation incorporates an additional potential term that represents the global image data likelihood. The intuition is that instead of just fitting the models locally on the current segmented subregions, we also seek for the best global description of the whole image, which results in more accurate model estimations and, thus, better OOI extraction.

The local-global energy minimization involves two steps: fixing the models, optimizing the OOI boundary by level set, and fixing the boundary curve, estimating the OOI and background models by fixed-point iterations. The fixed-point iterations, called quasi-semisupervised EM, is a robust method for estimating GMMs when some unknown portion of the data are labeled incorrectly. The added global likelihood potential increases the robustness.

Related work is summarized in Section 2. Our local-global energy formulation is presented in Section 3. In Section 4, we describe the energy minimization algorithm. Extensive experimental results are presented in Section 5. Finally, we conclude in Section 6.

- *G. Hua is with Microsoft Live Labs, One Microsoft Way, Redmond, WA 98052. E-mail: ganghua@microsoft.com.*
- *Z. Liu and Z. Zhang are with the Multimedia Collaboration Group, Microsoft Research, One Microsoft Way, Redmond, WA 98052. E-mail: {zliu, zhang}@microsoft.com.*
- *Y. Wu is with the Department of Electrical Engineering and Computer Science, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208. E-mail: yingwu@ece.northwestern.edu.*

Manuscript received 29 Sept. 2005; revised 15 Mar. 2006; accepted 27 Mar. 2006; published online 11 Aug. 2006.
Recommended for acceptance by R. Zabih.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0525-0905.

2 Related Work

Image segmentation by variational energy minimization can be traced back to SNAKES [6]. Later works include the Mumford-Shah model [7], [8], active contour with balloon forces [9], region competition approach [1], geodesic active contours [10], [11], region-based active contour [12], [13], [14], [15], geodesic active region [16], [2], [5], active contour with shape derivatives [17], etc. In practice, energy formulation purely based on image gradient [6], [9], [10], [8], [11] are vulnerable to a local solution, while using various image cues such as the intensity, color, and texture [1], [13], [2], [18], [14], [17], [15] can largely overcome this problem.

Region-based energy formulation can be categorized into two: *supervised* method [16], [2], [12], [18], [17] and *unsupervised* method [1], [5], [13], [14], [15]. Supervised methods assume the region models be known, while unsupervised methods need to jointly perform the segmentation and estimate the region models, which are normally solved by minimizing an energy with regard to the region boundary and region models alternatively. The methods of minimizing the energy with regard to the region boundary has evolved from the finite difference method (FDM) [6], [1] and finite element method (FEM) [9] to the level set method [19], [20], [10], [13], [2], [14], [17], [15].

Our local-global energy formulation combines different image cues including gradient, color, and spatial coherence of the pixels. It is different from previous works because of the additional global image likelihood potential.

3 Robust Variational Energy Formulation

For segmentation, it is essential to define the "coherence" of different image regions to group the pixels. It is natural to model it probabilistically. For general OOI extraction, it is not realistic to assume either the OOI or the background model to be a single Gaussian. In our formulation, we use GMM to model the OOI, denoted by $\mathcal{F}$, and the background, denoted by $\mathcal{B}$. That is, for $\mathcal{M} \in \{\mathcal{F}, \mathcal{B}\}$

$$P_{\mathcal{M}}(\mathbf{u}(x,y)) = P(\mathbf{u}(x,y)|(x,y) \in \mathcal{M}) = \sum_{i=1}^{K_{\mathcal{M}}} \pi_i^{\mathcal{M}} \mathcal{N}\left(\mathbf{u}(x,y)|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}}\right), \tag{1}$$

where π_i, $\vec{\mu}_i$, and $\vec{\Sigma}_i$ are, respectively, the weight, mean, and covariance of the ith mixture component, K is the number of mixtures, and $\mathbf{u}(x,y)$ is the feature vector at pixel (x,y). Denote the image data $\mathcal{I} = \mathcal{F} \cup \mathcal{B}$. Assuming image pixels are drawn *i.i.d.* from the two GMMs, the image data likelihood model is simply a mixture of the two, i.e.,

$$P_{\mathcal{I}}(\mathbf{u}(x,y)) = \omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}(x,y)) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u}(x,y)), \tag{2}$$

where $\omega_{\mathcal{F}} = P((x,y) \in \mathcal{F})$, $\omega_{\mathcal{B}} = P((x,y) \in \mathcal{B})$ are the priors such that $\omega_{\mathcal{F}} + \omega_{\mathcal{B}} = 1$.

3.1 Local Region Potential

Denote $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ as the estimates of OOI and background regions, where $\mathcal{I} = \mathcal{A}_{\mathcal{F}} \cup \mathcal{A}_{\mathcal{B}}$. The estimation quality can be evaluated by the local region likelihoods [1], [2], [16], i.e.,

$$\begin{aligned}\mathbf{E}_{hl} &= \prod_{(x,y)\in\mathcal{A}_\mathcal{F}} P(\mathbf{u}(x,y),(x,y)\in\mathcal{F}) \prod_{(x,y)\in\mathcal{A}_\mathcal{B}} P(\mathbf{u}(x,y),(x,y)\in\mathcal{B}) \\ &= \prod_{(x,y)\in\mathcal{A}_\mathcal{F}} \omega_\mathcal{F} P_\mathcal{F}(\mathbf{u}(x,y)) \prod_{(x,y)\in\mathcal{A}_\mathcal{B}} \omega_\mathcal{B} P_\mathcal{B}(\mathbf{u}(x,y)).\end{aligned} \tag{3}$$

Taking the logarithm on both sides, we obtain the local region likelihood potential, i.e.,

$$\mathbf{E}_h = \int_{\mathcal{A}_\mathcal{F}} \{\log P_\mathcal{F}(\mathbf{u}(x,y)) + \log\omega_\mathcal{F}\} + \int_{\mathcal{A}_\mathcal{B}} \{\log P_\mathcal{B}(\mathbf{u}(x,y)) + \log\omega_\mathcal{B}\}. \tag{4}$$

Our local region potential is more general than those in [1], [2], [16] since we incorporate the priors $\omega_\mathcal{F}$ and $\omega_\mathcal{B}$ in it. When we have no a priori knowledge about $\omega_\mathcal{F}$ and $\omega_\mathcal{B}$ (i.e., they are equal), (4) boils down to what is used in [1], [2], [16].

3.2 Global Image Data Likelihood Potential

Notice that (4) only locally evaluates the fitness of $\mathcal{A}_\mathcal{F}$ and $\mathcal{A}_\mathcal{B}$. When they are close to the ground truth, maximizing (4) gives the maximum likelihood estimation for the OOI and background models. But, in practice, the initial estimates of $\mathcal{A}_\mathcal{F}$ and $\mathcal{A}_\mathcal{B}$ are usually quite different from the ground truth. $\mathcal{A}_\mathcal{F}$ may not contain all the OOI pixels and it may even contain some background pixels. The same problem exists with $\mathcal{A}_\mathcal{B}$. This affects the accuracy of the estimates of the model parameters, which, in turn, affects the subsequent segmentation. This problem arises because we ignore the optimality of the global description of the entire image if we only maximize (4). Therefore, we propose to add a global image likelihood potential describing the entire image data. The intuition is that seeking for an optimal global description at the same time may largely reduce the negative effects of those erroneously labeled pixels.

The global image data likelihood is

$$\begin{aligned}\mathbf{E}_{ll} &= \prod_{(x,y)\in\mathcal{A}_\mathcal{F}\cup\mathcal{A}_\mathcal{B}} P_\mathcal{I}(\mathbf{u}(x,y)) \\ &= \prod_{(x,y)\in\mathcal{I}} \omega_\mathcal{F} P_\mathcal{F}(\mathbf{u}(x,y)) + \omega_\mathcal{B} P_\mathcal{B}(\mathbf{u}(x,y)).\end{aligned} \tag{5}$$

Taking the logarithm on both sides, the image data likelihood potential is defined as

$$\mathbf{E}_l = \int_\mathcal{I} \log\{\omega_\mathcal{F} P_\mathcal{F}(\mathbf{u}(x,y)) + \omega_\mathcal{B} P_\mathcal{B}(\mathbf{u}(x,y))\}. \tag{6}$$

Later in the experiments, we will demonstrate that incorporating it does significantly reduce the negative effects of the erroneously labeled pixels in the model estimation step.

3.3 Boundary Potential

Image edges provide valuable information for segmentation. A lot of work incorporates a boundary edge potential in a variational energy formulation [6], [9], [2]. Denoting $\Gamma(c): c\in[0,1]\to(x,y)\in\mathbf{R}^2$ as the closed curve between $\mathcal{A}_\mathcal{F}$ and $\mathcal{A}_\mathcal{B}$ such that $\Gamma(c)=\mathcal{A}_\mathcal{F}\cap\mathcal{A}_\mathcal{B}$, we adopt a geodesic boundary potential to evaluate the alignment of $\Gamma(c)$ to image edges, i.e.,

$$\begin{aligned}\mathbf{E}_e(\Gamma(c)) &= \int_0^1 \frac{1}{1+|\mathbf{g}_x(\Gamma(c))|+|\mathbf{g}_y(\Gamma(c))|}|\dot\Gamma(c)|dc \\ &= \int_0^1 G(\Gamma(c))|\dot\Gamma(c)|dc,\end{aligned} \tag{7}$$

where $(\mathbf{g}_x,\mathbf{g}_y)$ is the image gradient vector, and $\dot\Gamma(c)$ is the derivative of $\Gamma(c)$. Minimizing $\mathbf{E}_e$ will align $\Gamma(c)$ to the image pixels with maximum gradient while $\dot\Gamma(c)$ ensures the smoothness.

3.4 Boundary, Region, and Data Likelihood Synergism

Our energy functional is then defined as the synergism of the above three potentials, i.e.,

$$\mathbf{E}_p(\Gamma(c),P_\mathcal{I}) = \alpha\mathbf{E}_e - \beta\mathbf{E}_h - \gamma\mathbf{E}_l, \tag{8}$$

where α, β, and γ are positive, which balance the three potentials and $\alpha+\beta+\gamma=1$.

4 Energy Minimization Algorithms

The energy is minimized iteratively. Each iteration consists of two steps: First, we fix $P_\mathcal{I}(\mathbf{u})$ and solve for $\Gamma(c)$ by level set. Second, we fix $\Gamma(c)$ and re-estimate $P_\mathcal{I}(\mathbf{u})$ by fixed-point iteration.

4.1 Boundary Optimization by Level Set

At the first step, we fix $P_\mathcal{I}(\mathbf{u})$ and minimize $\mathbf{E}_p$ with regard to $\Gamma(c)$. This is achieved by gradient decent. Taking the variation of $\mathbf{E}_p(\Gamma(c),P_\mathcal{I})$ with regard to $\Gamma(c)$, we have

$$\begin{aligned}\frac{\partial\mathbf{E}_p}{\partial\Gamma(c)} = \Bigg\{&\beta\log\left[\frac{\omega_\mathcal{F}P_\mathcal{F}(\mathbf{u}(\Gamma(c)))}{\omega_\mathcal{B}P_\mathcal{B}(\mathbf{u}(\Gamma(c)))}\right] \\ &+\alpha\Big[G(\Gamma(c))\mathcal{K}(\Gamma(c)) - \nabla G(\Gamma(c))\cdot\vec{n}(\Gamma(c))\Big]\Bigg\}\cdot\vec{n}(\Gamma(c)),\end{aligned} \tag{9}$$

where $\vec{n}(\cdot)$ is the normal vector of $\Gamma(c)$ pointing outward and $\mathcal{K}(\cdot)$ is the curvature. One interesting observation is that the partial variation in (9) is very similar to that in [2], [16] except that it contains the additional parameters $\omega_\mathcal{F}$ and $\omega_\mathcal{B}$. This is easy to understand because the global image likelihood potential $\mathbf{E}_l$ does not rely on the boundary $\Gamma(c)$.

We use the level set technique to solve the above PDE. At each step t during the curve optimization, $\Gamma(c,t)$ is represented by the zero level set of a two-dimensional surface $\varphi(x,y,t)$ (e.g., a signed distance function), i.e., $\Gamma(c,t):=\{(x,y)|\varphi(x,y,t)=0\}$. Then, we have

$$\begin{aligned}\frac{\partial\varphi(x,y,t)}{\partial t} = \Bigg\{&\beta\log\left[\frac{\omega_\mathcal{F}P_\mathcal{F}(\mathbf{u}(x,y))}{\omega_\mathcal{B}P_\mathcal{B}(\mathbf{u}(x,y))}\right] \\ &+\alpha\left[G(x,y)\mathcal{K}(x,y) - \nabla G(x,y)\cdot\frac{\nabla\varphi}{|\nabla\varphi|}\right]\Bigg\}|\nabla\varphi|,\end{aligned} \tag{10}$$

where

$$\mathcal{K}(x,y) = \frac{\varphi_{xx}\varphi_y^2 - 2\varphi_{xy}\varphi_x\varphi_y + \varphi_{yy}\varphi_x^2}{(\varphi_x^2+\varphi_y^2)^{\frac{3}{2}}}$$

among which φ_x and φ_y, and φ_{xx}, φ_{yy}, and φ_{xy} are the first and second order partial derivatives of $\varphi(\cdot)$. The surface evolution is then calculated by

$$\begin{aligned}\varphi(x,y,t+\tau) = \varphi(x,y,t) + \tau\cdot\Bigg\{&\beta\log\left[\frac{\omega_\mathcal{F}P_\mathcal{F}(\mathbf{u}(x,y))}{\omega_\mathcal{B}P_\mathcal{B}(\mathbf{u}(x,y))}\right]|\nabla\varphi(\cdot)| \\ &+\alpha\left[G(x,y)\mathcal{K}(x,y) - \nabla G(x,y)\cdot\frac{\nabla\varphi(\cdot)}{|\nabla\varphi(\cdot)|}\right]|\nabla\varphi(\cdot)|\Bigg\},\end{aligned} \tag{11}$$

where τ is the step size. We have $\Gamma(c,t+\tau)=\{(x,y)|\varphi(x,y,t+\tau)=0\}$. In practice, all derivatives are replaced by discrete differences, i.e., φ_t is approximated by forward differences and φ_x and φ_y are approximated by central differences.

4.2 Image Data Model Estimation

At the second step, we fix $\Gamma(c)$ and minimize $\mathbf{E}_p$ with regard to $P_\mathcal{I}(\mathbf{u})$. This involves the minimization of $\mathbf{E}_p$ with regard to $\boldsymbol{\Theta}=\left\{\omega_\mathcal{F},\omega_\mathcal{B},\{\pi_i^\mathcal{F},\vec{\mu}_i^\mathcal{F},\vec{\Sigma}_i^\mathcal{F}\}_{i=1}^{K_\mathcal{F}},\{\pi_i^\mathcal{B},\vec{\mu}_i^\mathcal{B},\vec{\Sigma}_i^\mathcal{B}\}_{i=1}^{K_\mathcal{B}}\right\}$, the parameter set of $P_\mathcal{I}(\mathbf{u})$. Notice that $\mathbf{E}_e$ is independent of $\boldsymbol{\Theta}$. By taking the derivatives of $\mathbf{E}_p$ with regard to all parameters and setting them to zero, after easy manipulations, we obtain the following fixed-point equations: For $\mathcal{M}\in\{\mathcal{F},\mathcal{B}\}$

TABLE 1
Comparison of Local-Global Energy Minimization (LGEM) and Local Energy Minimization (LEM)

	$\beta = 0.01$	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.15$	$\beta = 0.2$	$\beta = 0.25$	$\beta = 0.3$	$\beta = 0.5$
LEM	5.6 ± 3.1	5.2 ± 2.8	5.5 ± 3.0	5.4 ± 3.0	5.4 ± 3.0	5.5 ± 3.0	5.7 ± 3.2	5.4 ± 2.9
LGEM	0.2 ± 0.2	0.5 ± 0.6	1.7 ± 1.4	3.5 ± 2.6	4.8 ± 3.2	5.4 ± 3.3	5.7 ± 3.5	5.6 ± 3.3
Better LGEM (%)	100%	99.9%	99.3%	87.6%	72.1%	57.2%	49.6%	34.2%

The table shows the average joint KLs distance and its standard deviation between the estimated model and the ground truth from 1,000 simulations for each β.

$$\omega_{\mathcal{M}}^{*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{M}}} 1 + \gamma \int_{\mathcal{I}} \frac{2\omega_{\mathcal{M}} P_{\mathcal{M}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\gamma \int_{\mathcal{I}} \frac{P_{\mathcal{M}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}, \tag{12}$$

$$\pi_i^{\mathcal{M}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{M}}} \frac{\pi_i^{\mathcal{M}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{M}} P_{\mathcal{M}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{M}}} \frac{2\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{M}} P_{\mathcal{M}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}, \tag{13}$$

$$\vec{\mu}_i^{\mathcal{M}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{M}}} \frac{\mathbf{u}\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{M}} P_{\mathcal{M}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathbf{u}\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{M}}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{M}} P_{\mathcal{M}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}, \tag{14}$$

$$\vec{\Sigma}_i^{\mathcal{M}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{M}}} \frac{(\mathbf{u}-\vec{\mu}_i^{\mathcal{M}})(\mathbf{u}-\vec{\mu}_i^{\mathcal{M}})^T \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{M}} P_{\mathcal{M}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{(\mathbf{u}-\vec{\mu}_i^{\mathcal{M}})(\mathbf{u}-\vec{\mu}_i^{\mathcal{M}})^T \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{M}}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{M}}, \vec{\Sigma}_i^{\mathcal{M}})}{\omega_{\mathcal{M}} P_{\mathcal{M}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \vec{\Sigma}_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}. \tag{15}$$

During the fixed-point iterations, $\omega_{\mathcal{F}}^{*}$, $\omega_{\mathcal{B}}^{*}$, $\pi_i^{\mathcal{F}*}$, and $\pi_i^{\mathcal{B}*}$ are normalized so that $\omega_{\mathcal{F}}^{*} + \omega_{\mathcal{B}}^{*} = 1, \sum_{i=1}^{K_{\mathcal{F}}} \pi_i^{\mathcal{F}*} = 1, \sum_{i=1}^{K_{\mathcal{B}}} \pi_i^{\mathcal{B}*} = 1$. This set of fixed-point equations, called quasi-semisupervised EM, is a robust scheme for estimating the GMMs of two classes in the case when all the data are labeled, but some unknown portion of the labels are erroneous. Here, the OOI and background are the two classes and $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ are the inaccurate labels of the OOI and background pixels.

It is easy to figure out that estimating the model parameters by optimizing the local region potential $\mathbf{E}_h$ is purely a supervised estimation, while estimating the model parameters by optimizing the global likelihood potential $\mathbf{E}_l$ is purely an unsupervised estimation. In the iterative energy minimization process, purely supervised estimation is confronted by the erroneously labeled pixels, while purely unsupervised estimation does not have this problem but it totally ignores the useful information from those correctly labeled pixels. In [21], a robust method of estimating Fisher discriminant under the presence of label noise is presented, but it restricts to Gaussian distributions which may not be that interesting in our case of estimation GMMs.

The fixed-point equations derived above indeed seek a tradeoff between the supervised estimation and the unsupervised estimation. This can be easily observed in the numerator of (14), where the first integral over $\mathcal{A}_{\mathcal{M}}$ is the estimation from the inaccurately labeled data while the second integral over $\mathcal{I}$ is a soft classification of the image pixels by the current estimation of the image data model. Those image pixels which have been labeled to be in $\mathcal{A}_{\mathcal{M}}$ and which have also been classified with high confidence as $\mathcal{M}$ will be assigned with more weights. This suppresses the negative effects of those erroneously labeled data.

5 Experiments

We first use a synthetic example to demonstrate that the additional global likelihood potential does improve the accuracy of the model estimation. Since the model estimation is independent of the boundary potential, the synthetic example does not include the boundary energy term. We then present extensive experiments on real data to automatically extract OOI from images.

5.1 Validation of Minimizing the Local-Global Energy

In this experiment, the ground-truth of the one-dimensional data model is

$$\begin{aligned} P(\mathbf{d}|\boldsymbol{\Theta}) &= \omega_1 P_1(\mathbf{d}|\boldsymbol{\Theta}_1) + \omega_2 P_2(\mathbf{d}|\boldsymbol{\Theta}_2) \\ &= \omega_1 \big(\pi_{11}\mathcal{N}(\mathbf{d}|\mu_{11}, \sigma_{11}^2) + \pi_{12}\mathcal{N}(\mathbf{d}|\mu_{12}, \sigma_{12}^2)\big) \\ &+ \omega_2 \big(\pi_{21}\mathcal{N}(\mathbf{d}|\mu_{21}, \sigma_{21}^2) + \pi_{22}\mathcal{N}(\mathbf{d}|\mu_{22}, \sigma_{22}^2)\big), \end{aligned} \tag{16}$$

where $\boldsymbol{\Theta}$ is the parameters set. We denote $\boldsymbol{\omega}$ as a binomial random variable with p.m.f. $\{\omega_1, \omega_2\}$.

With a specific $\boldsymbol{\Theta}$, we randomly draw a set $\mathcal{D}$ of 20,000 data samples and record the set $\mathcal{L}$ of ground-truth labels, which indicates whether a sample is from P_1 or P_2. We denote $\mathcal{L}_1 = \{l_i = 1\}$ and $\mathcal{L}_2 = \{l_i = 2\}$. To simulate the inaccurate labeling, we randomly exchange 30 percent labels between $\mathcal{L}_1$ and $\mathcal{L}_2$. We denote the exchanged label set as $\mathcal{Z}_1$ and $\mathcal{Z}_2$, which are regarded as the known conditions for model estimation. We then compare the model estimated by minimizing the local energy $-\mathbf{E}_h$ and the model estimated by minimizing the local-global energy $-\beta\mathbf{E}_h - (1-\beta)\mathbf{E}_l$. In the experiments, the local-global energy minimization (LGEM) is performed by the quasi-semisupervised EM algorithm similar to that in Section 4.2. The local energy minimization (LEM) is performed by applying the classical EM algorithm [22] independently to the two data sets induced by $\mathcal{Z}_1$ and $\mathcal{Z}_2$.

Denote $P^{*}(\mathbf{d}) = \omega_1^{*} P_1^{*}(\mathbf{d}) + \omega_2^{*} P_2^{*}(\mathbf{d})$ as the estimated distribution and ω^{*} as the binomial random variable with p.m.f. $\{\omega_1^{*}, \omega_2^{*}\}$. We then evaluate the quality of the estimated distribution with regard to the ground truth by the following *joint KLs distance*, i.e.,

$$\mathcal{D}(P^{*}, P) = KL_s(\omega^{*}, \omega) + KL_s(P_1^{*}, P_1) + KL_s(P_2^{*}, P_2), \tag{17}$$

where $KL_s(f, g) = \frac{KL(f\|g) + KL(g\|f)}{2}$ is the symmetric KL distance. Notice that by definition, when the joint KLs distance in (17) is small, we can assure that all the estimated parameters are close to the ground truth.

We have extensively evaluated the quality of the estimated models from both algorithms. Fixing a β, we randomly generate 1,000 data models and, thus, run 1,000 simulations of the experiments described above. For both algorithms, in each simulation we randomly choose 10 different initializations and the best results are adopted. The experimental results are listed in Table 1. The third row of Table 1 presents the percentage of the 1,000 simulations for a fixed β in which the LGEM estimated better models than the LEM.

Table 1 clearly shows that with 30 percent erroneous labels, the estimated models from LGEM is significantly more accurate than those from LEM when we set β be 0.01 for the local-global energy, i.e., the average $\mathcal{D}$ for LGEM is only 0.2 with standard deviation 0.2 over the 1,000 simulations. While the LEM obtains a average $\mathcal{D}$ of 5.6 with standard deviation 3.1. We can also notice that if we increase β, the models estimated by LGEM will degrade. When $\beta > 0.2$, the performance of LGEM is almost the same or even worse than LEM.

We observed similar results for 20 percent and 10 percent erroneous labels, although β needs to be larger to degrade the

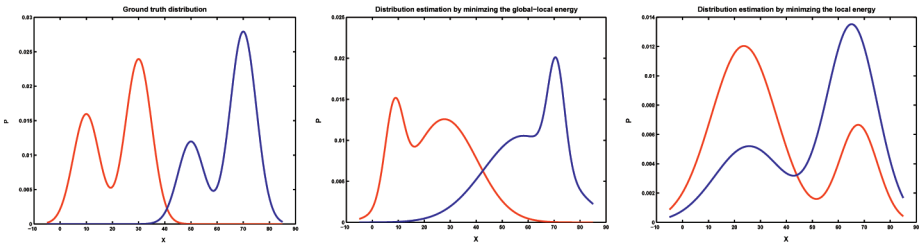

Fig. 1. Visual comparison of model estimation by local and local-global energy minimization in one simulation ($\beta = 0.05$). (a) Ground truth. (b) Local-global energy estimation. (c) Local energy estimation.

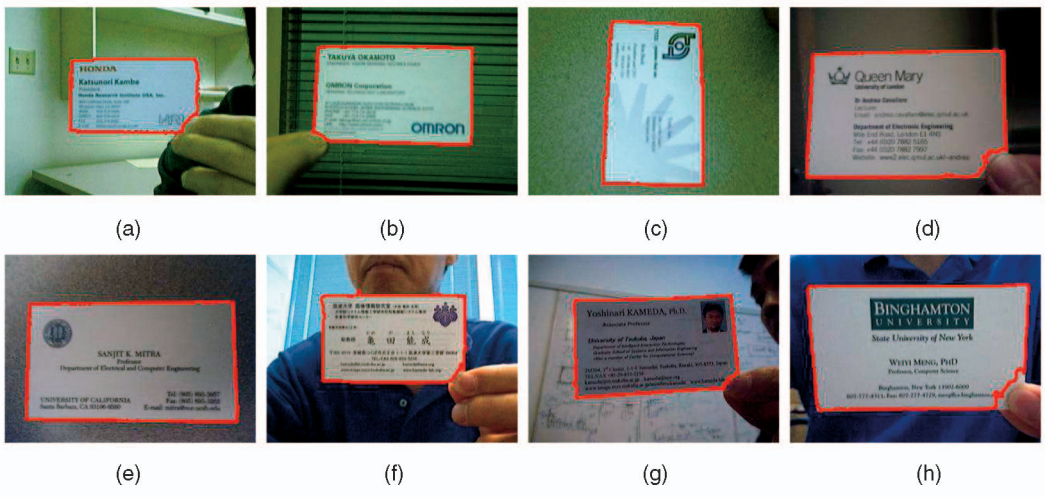

Fig. 2. Card segmentation results.

performance of LGEM to that of LEM. All these suggested us to set a small β for the local-global energy. We usually set it to be 0.01 since we also found that setting $\beta < 0.01$ will not improve the performance too much while the quasi-semisupervised EM may take much longer to converge. Following the idea of [23], theoretic analysis of choosing β may be possible, we defer it to our future work. As pointed out in [24], estimating the GMM in a purely unsupervised fashion can hardly keep the identity of the Gaussian component. We do not have this problem because we combine the global energy with the local region energy. We plotted in Fig. 1 the models estimated from LGEM, from LEM, and the ground truth in one simulation when $\beta = 0.05$. The model from LGEM is far more accurate than that from LEM.

5.2 Automatic Extraction of Objects of Interest

The weak assumption of focus-of-attention enables us to build a fully automatic system to extract the OOI from images. Some implementation details are as follows:

- *Feature*: $\mathbf{u} = (L, U, V, x, y)$, i.e., the LUV pixel values with the image coordinates [25].
- *Model*: $K_{\mathcal{F}} = 2$ and $K_{\mathcal{B}} = 8$.
- *Surface Initialization*: The level set surface is initialized from a centered rectangle with $\frac{1}{8}$ image width and length by a signed distance transform.
- *Foreground initialization*: We sort the pixels inside the initial rectangle according to their L value. We take the two average feature vectors of the lightest 10 percent pixels and the darkest 10 percent pixels inside it as the seeds for the mean-shift to obtain two feature modes. The two modes are adopted to initialize $\vec{\mu}_1^{\mathcal{F}}$ and $\vec{\mu}_2^{\mathcal{F}}$. The $\pi_1^{\mathcal{F}}$ and $\pi_2^{\mathcal{F}}$ are initialized as 0.5. Each $\vec{\Sigma}_i^{\mathcal{F}}$ is initialized with the same diagonal matrix: the variance of (x, y) are set to be the square of the $\frac{1}{5}$ of the image width and height; the variances of (L, U, V) are all initialized as 25.
- *Background initialization*: The average feature vectors inside eight 10×10 rectangles around the image borders are used as the seeds of mean-shift to obtain $K_{\mathcal{B}} = 8$ feature modes. The modes are used to initialize the $\vec{\mu}_i^{\mathcal{B}}$ for $i = 1, \ldots, 8$. Each $\vec{\Sigma}_i^{\mathcal{B}}$ has the same initialization as the $\vec{\Sigma}_j^{\mathcal{F}}$. All the $\pi_i^{\mathcal{B}}$s are initialized as $\frac{1}{8}$.
- INITIALIZATION OF $\omega_{\mathcal{F}}$ AND $\omega_{\mathcal{B}}$: They are initialized as $\frac{1}{2}$.
- *Convergence criterion*: When the OOI region has less than 1 percent change in two consecutive iterations, we consider that the algorithm is converged.

We now present the OOI extraction results on images including business card, road signs, and other more general objects.

5.2.1 Business Card Extraction

We first tested our OOI extraction algorithm on a set of business card images. One interesting application is in mobile note taking. One can use his/her mobile phone camera to scan and thus manage the business card he received from the others. The proposed approach generally produces satisfactory results and we achieve 95 percent successful rate on over 300 images tested. We regard a result as being successful if it almost matches the human segmentation. The evaluation was done subjectively. Fig. 2 presents some of the result images.

5.2.2 Segmentation of Road Sign Images

We have collected a set of 37 road sign images in which the road signs are at the focus of attention. It contains road signs of different shapes and different poses with a large variety of backgrounds. We asked seven people to evaluate the quality of the extraction results by giving a rating of "good," "fair," or "bad" to

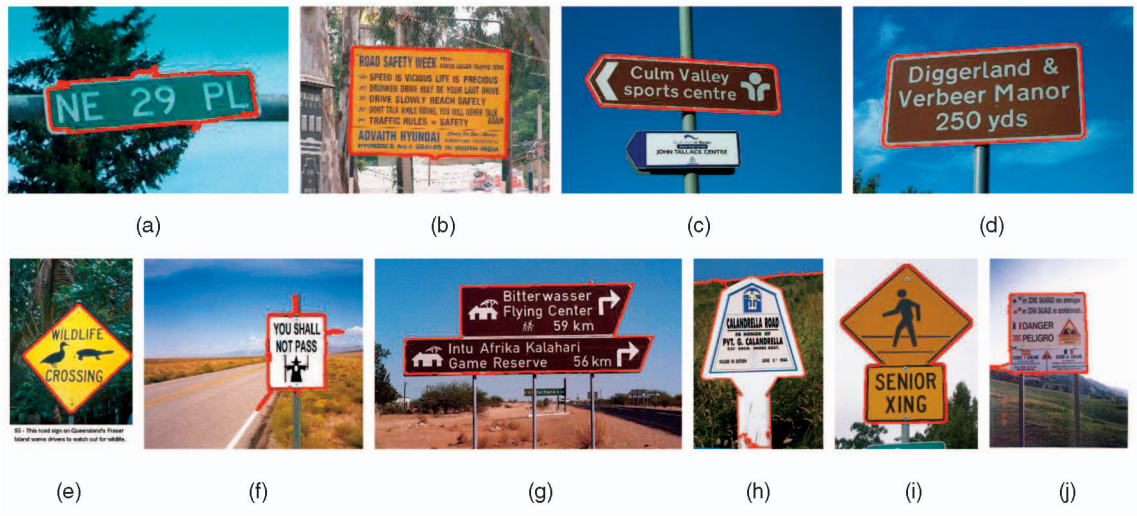

Fig. 3. Segmentation results of road sign images.

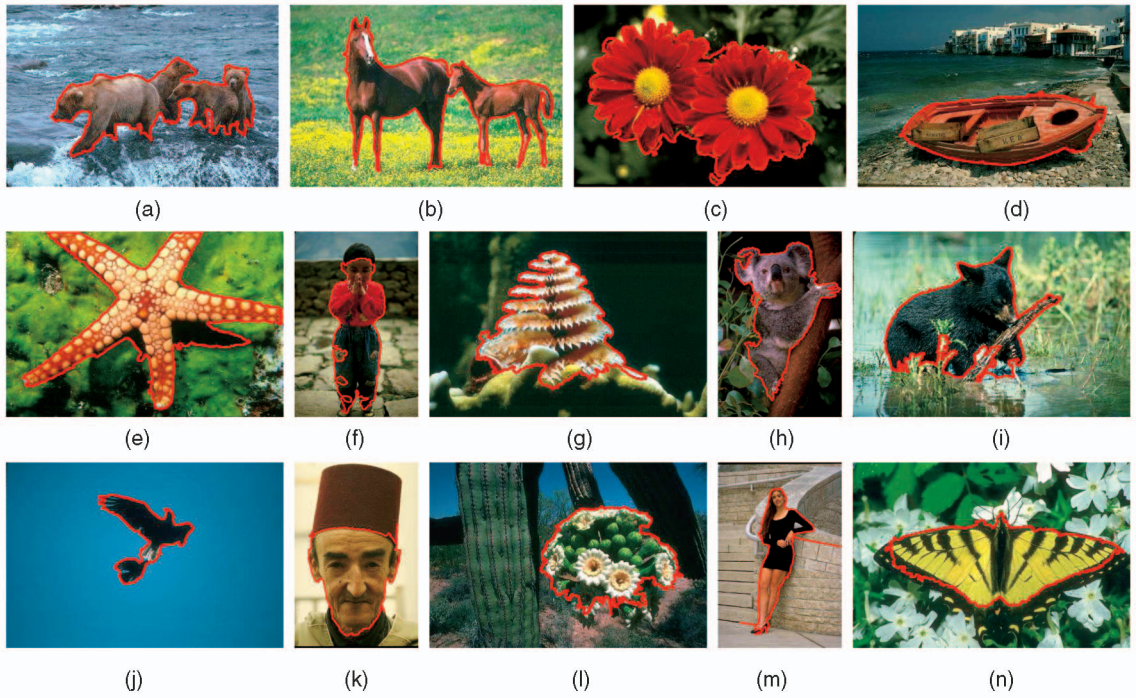

Fig. 4. Extract OOIs on Berkeley image database.

each image. The majority rule is adopted to categorize each result. Overall there are 27 good, 5 fair, and 5 bad. Some of the successful results are shown in Fig. 3.

Two reasons may cause the unsatisfactory results: 1) The OOIs are too small or too thin in the images. In that case, the initial OOI region may contain a large number of background pixels. 2) There are very strong spurious edges surrounding the OOI while there is not enough contrast between the foreground and the background colors to overcome the biased energy force from the spurious edges. One possible solution might be to reduce α, but how to tune it adaptively is an open issue. Note that these reasons also apply to the bad examples to be presented next.

5.2.3 Segmentation of General Objects

To test the effectiveness and robustness of the proposed algorithm for extracting general OOIs, we have tested on a set of 63 images, in which the OOIs are at the focus of attention, from the Berkley image database [26]. This set of images are more challenging. With the same evaluation method, 31 extraction results are good, 15 are fair, and 17 are bad. We present some typical successful results in Fig. 4.

5.2.4 Comparison with the Energy Formulation without the Global Potential

For comparison, we also implemented the algorithm with the energy formulation without the global likelihood potential. Under this formulation, in the model estimation step, the classical EM algorithm [22] is applied to $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ independently to obtain the OOI and background GMMs. Its performance is, in general, inferior to the proposed approach. We tested it on the 37 road sign images and 63 images from the Berkeley image database. The comparison results[1] are summarized in Table 2. As we can notice, for the 37 road sign images, the local energy formulation produces 19 good, 7 fair, and 8 bad results, which are significantly inferior to

1. The image results can be accessed at http://www.ece.northwestern.edu/~ganghua/PAMI2006/.

TABLE 2
Comparison Results of the Proposed Local-Global Variational Energy Formulation and the Variational Energy Formulation without the Global Potential Term on the Road Sign Images and Berkeley Images

	Local-global Formulation			Local Formulation		
	Good	Fair	Bad	Good	Fair	Bad
Road Sign Images	27	5	5	19	7	8
Berkeley Images	31	15	17	20	16	27
Overall	58	20	22	38	23	36

the 27 good, 5 fair, and 5 bad results obtained with our local-global energy formulation. For the 63 Berkeley images, the local energy formulation produces 19 good, 16 fair, and 28 bad results, which are again significantly inferior to the results obtained with our approach.

For one-on-one comparison on each image, we also found that the extraction results from our local-global energy formulation is always superior to those from the local energy formulation. In other words, whenever the local energy formulation produces a good result, our approach can also produce a good result if not better. On the other hand, on a significant number of test images, our approach produced good results while the local energy formulation failed. It is the global image likelihood potential that makes this difference. It enables the model estimation step to be more accurate.

6 Conclusion and Future Work

We have proposed a novel local-global variational energy formulation to automatically extract OOI from images, and developed an efficient iterative scheme to minimize it. Our main contributions are: 1) the incorporation of a global image likelihood potential for better estimating the OOI and background models and 2) a set of fixed-point equations which we call quasi-semisupervised EM for robust estimation of GMMs from inaccurately labeled data. Extensive experiments demonstrated the efficacy of our approach. Future work includes extending the variational energy formulation for automatic extraction of multiple objects.

Acknowledgments

The authors thank the reviewers for their constructive suggestions. The majority of the work was carried out when Gang Hua was a summer intern at Microsoft Research, Redmond, Washington.

References

[1] S.C. Zhu and A. Yuille, "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation," *IEEE Trans. Pattern Recognition and Machine Intelligence,* vol. 18, no. 9, pp. 884-900, Sept. 1996.
[2] N. Paragios and R. Deriche, "Geodesic Active Regions and Level Set Methods for Supervised Texture Segmentation," *Int'l J. Computer Vision,* pp. 223-247, 2002.
[3] A. Blake, C. Rother, M. Brown, P. Pérez, and P. Torr, "Interactive Image Segmentation Using an Adaptive Gaussian Mixture MRF Model," *Proc. Eighth European Conf. Computer Vision,* pp. 428-441, 2004.
[4] C. Rother, V. Kolmogorov, and A. Blake, "'Grabcut'—Interactive Foreground Extraction Using Iterated Graph Cuts," *ACM Trans. Graphics (Proc. SIGGRAPH '04),* pp. 309-314, 2004.
[5] M. Rousson, T. Brox, and R. Deriche, "Active Unsupervised Texture Segmentation on a Diffusion Based Feature Space," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 699-704, June 2003.
[6] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *Int'l J. Computer Vision,* vol. 1, pp. 321-331, 1987.
[7] D. Mumford and J. Shah, "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problem," *Comm. Pure and Applied Math.,* vol. 42, pp. 577-584, 1989.
[8] A. Tsai, J.A. Yezzi, and A.S. Willsky, "A Curve Evolution Approach to Smoothing and Segmentation Using the Mumford-Shah Functional," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1119-1124, June 2000.
[9] L.D. Cohen and I. Cohen, "Finite-Element Methods for Active Contour Models and Balloons for 2-D and 3-D Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 11, pp. 1131-1147, Nov. 1993.
[10] V. Casselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *Int'l J. Computer Vision,* vol. 22, no. 1, pp. 61-79, 1997.
[11] N. Paragios, O. Mellina Gottardo, and V. Ramesh, "Gradient Vector Flow Fast Geodesic Active Contours," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 67-73, July 2001.
[12] J.A. Yezzi, A. Tsai, and A.S. Willsky, "A Statistical Approach to Snakes for Bimodal and Trimodal Imagery," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 898-903, Sept. 1999.
[13] T.F. Chan and L.A. Vese, "Active Contours without Edges," *IEEE Trans. Image Processing,* vol. 10, no. 2, pp. 266-277, Feb. 2001.
[14] S. Jehan-Besson, M. Barlaud, and G. Aubert, "Video Object Segmentation Using Eulerian Region-Based Active Contours," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 353-361, July 2001.
[15] J. Kim, J.W. Fisher III, A.J. Yezzi, M. Çetin, and A.S. Willsky, "A Nonparametric Statistical Method for Image Segmentation Using Information Theory and Curve Evolution," *IEEE Trans. Image Processing,* vol. 14, no. 10, pp. 1486-1502, Oct. 2005.
[16] N. Paragios and R. Deriche, "Geodesic Active Contours for Supervised Texture Segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 1034-1040, 1999.
[17] S. Jehan-Besson, M. Barlaud, and G. Aubert, "Shape Gradients for Histogram Segmentation Using Active Contours," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 408-415, Oct. 2003.
[18] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia, "A Level Set Model for Image Classification," *Int'l J. Computer Vision,* vol. 40, no. 3, pp. 187-197, Mar. 2000.
[19] S. Osher and J.A. Sethian, "Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulation," *J. Computational Physics,* vol. 79, pp. 12-49, 1988.
[20] R. Malladi, J.A. Sethian, and B.C. Vemuri, "Shape Modeling with Front Propagation: A Level Set Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 2, pp. 158-175, Feb. 1995.
[21] N.D. Lawrence and B. Schölkopf, "Estimating a Kernel Fisher Discriminant in the Presence of Label Noise," *Proc. Int'l Conf. Machine Learning,* C. Brodley and A.P. Danyluk, eds., pp. 306-313, 2001.
[22] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc., Series B,* vol. 39, no. 1, pp. 1-38, 1977.
[23] A. Corduneanu and T. Jaakkola, "Continuation Methods for Mixing Heterogenous Sources," *Proc. Uncertainty in Artificial Intelligence Conf.,* pp. 111-118, 2002.
[24] X. Zhu, J. Yang, and A. Waibel, "Segmenting Hands of Arbitrary Color," *Proc. IEEE Int'l Conf. Automatic Face Recognition,* pp. 446-453, Mar. 2000.
[25] D. Comaniciu and P. Meer, "Mean-Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 1-18, May 2002.
[26] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," *Proc. Eighth IEEE Int'l Conf. Computer Vision,* vol. 2, pp. 416-423, July 2001.