Iterative local-global energy minimization for automatic extraction of object of interest

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

Gang Hua and Dr. Ying Wu are with the Department of Electrical Engineering and Computer Science, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208. Phone: 847-491-4466, Fax: 847-491-4455, E-mail:{ganghua, yingwu}@ece.northwestern.edu.

Dr. Zicheng Liu and Dr. Zhengyou Zhang are with the Speech Technology Group, Microsoft Research, One Microsoft Way, Redmond, WA 98052. E-mail: {zliu,zhang}@microsoft.com

Abstract

In this paper, we propose a novel variational energy formulation for image segmentation. Traditional variational energy formulation for image segmentation like that in [1] only incorporates local region potentials with a Gaussian distribution on each region. We argue that for segmentation of natural objects, Gaussian mixture model (GMM) needs to be adopted to capture the appearance variation of the objects. Moreover, we introduce a global image data likelihood potential to address the problem that each local region usually contains a portion of incorrectly classified pixels during the iterations. By combining it with local region potentials, we obtain more robust and accurate estimation of the foreground and background distributions. The minimization of the proposed local-global energy functional is achieved in two steps: the evolution of the foreground and background boundary curve by level set; and the robust estimation of the foreground and background model by fixed-point iteration, called quasi-semi-supervised EM, which is particularly suited for the learning problem where some unknown portion of the data are labeled incorrectly. Extensive experimental results including both business card extraction, road sign extraction and general object-of-interest segmentation, demonstrate the robustness, effectiveness, and efficiency of the proposed approach.

Index Terms

Variational energy, Level set, Fixed-point iteration, Model estimation, Semi-supervised learning

I. INTRODUCTION

Automatic extraction of objects of interest (OOI) from still images is an important problem of early vision with applications in object recognition, image painting, video content analysis, visual surveillance, etc. Given an arbitrary image, the object of interest is usually subjective, but it 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 system to extract the OOIs. Notice that this assumption does not tell us where the boundary of the OOI is.

In order to extract the OOI, it is desirable to have the foreground and background models. The extraction of the OOI and the estimation of foreground and background models is intrinsically a *chicken-and-egg* problem. If we have the *a priori* knowledge of the OOI model, we can directly separate the OOI from the background like what is done in the geodesic active region approach [2]. On the other hand, if we have already extracted the OOI from the image, we can

easily estimate the OOI and the background models from the segmented image data. Problems like this are usually solved in an iterative way. At each iteration, the current estimates of the OOI and background models are fixed first, and the segmentation is performed. Based on the segmentation, the OOI and background models are then re-estimated. These two steps can be formulated as an energy minimization problem such as the region competition approach [1] and the GrabCut system [3]. Both the region competition and the GrabCut techniques model coherent image regions in a probabilistic way. The region competition algorithm models each coherent image region as a Gaussian distribution on the pixel intensities. The GrabCut models both the foreground and background as Gaussian mixture models on the RGB channels of the pixels. Both algorithms iteratively minimize an energy function whose unknowns include the segmentation results and the subregion model parameters.

One problem with the iteration process is that at each iteration the model estimation for each subregion is based on inaccurately labeled image pixels since the initial partition is usually not perfect. The incorrectly labeled pixels will affect the accuracy of the model parameters which in turn will affect the subsequent segmentation. How to reduce the negative effect of the incorrectly labeled pixels is the central focus of this paper.

We propose a novel variational energy formulation for the problem of the OOI extraction, which combines different image cues including gradient, color distribution, and spatial coherence of the image pixels. Our energy formulation differentiates from previous works ([1], [3]) in that we incorporate a potential function that represents the global image data likelihood. The intuition of incorporating this term is that instead of just fitting the models locally for each subregion on the inaccurately labeled image pixels, we also want to seek for a global description of the whole image data in the energy minimization process.

The minimization of the proposed energy functional involves two steps: the optimization of the OOI and background boundary curve by level set with the model distributions fixed; and the robust estimation of the OOI and background models by a fixed-point iteration with the boundary curve fixed. The robustness of the model estimation results from incorporating the global image likelihood potential. What is more interesting is that the fixed-point iteration reveals a robust computational paradigm of model estimation for Gaussian mixtures when some unknown portion of the data are labeled incorrectly. This is different from semi-supervised learning because in semi-supervised learning, the labels are assumed to be correct. To the best of our knowledge, we

are the first to propose such a machine learning technique which we call quasi-semi-supervised EM.

The remainder of the paper is organized as follows: the related work will be summarized and discussed in Section II; the detailed discussion of our variational energy formulation with the global image likelihood potential is presented in Section III; in Section IV, we describe the details of the iterative minimization algorithm including the optimization of the boundary curve by level set, and the detailed derivation of the quasi-semi-supervised EM algorithm for the robust estimation of the OOI and background models; in Section V, extensive experimental results on business card scanning, road signs extraction and general OOI extraction are presented and discussed; finally we conclude and discuss future work in Section VI.

II. Related work

Image segmentation and foreground/background separation is a fundamental yet difficult problem in computer vision. There has been a lot of work in this area, and it is formidable to enumerate all of them. We will only mention a few that are most related to our work.

One popular approach is to formulate the segmentation problem as an energy minimization problem. This approach can be roughly categorized as two mainstreams: variational energy minimization which usually involves solving a partial differential equation (PDE), and graph energy minimization which minimizes an energy function by graph-cut.

The research of image segmentation by variational energy minimization can be traced back to the active contour SNAKES [4]. Later work include the Mumford-Shah model [5], the active contour with balloon forces [6], the region competition algorithm [1], the geodesic active contours [7], the active contour without edges [8], the geodesic active region [2], etc. The energy functionals constructed in this track are usually formulated on the region boundary curves [4], [6], [7] and/or over the regions partitioned by the boundary curves [1], [8], [2]. In practice, energy functionals based purely on image gradient information like what was proposed in [4], [6], [7] are easy to get stuck in a local optima especially when there are many spurious edges in the image. On the other hand, using the intensity, color and texture distributions [1], [8], [2] of the image pixels over the regions to formulate the energy functional can largely overcome this problem. In principle, we can obtain a better energy formulation by combining the edge information and the feature distribution of the image pixels [2]. The minimization of this type of

variational energy has evolved from the traditional finite difference method (FDM) [4], [1] and the finite element method (FEM) [6] to the more advanced level set method [9], [10], [7], [8], [2]. There are a lot of work on the implementation of the level set method to reduce the computation involved during the evolution of the implicit level set surface, such as the narrow-band level set method [11], the level set without re-initialization [12] and the fast level set implementation without solving PDEs [13]. In fact, all these efficient level set algorithms take advantage of the property of the signed distance function [14], which is usually adopted as the implicit level set surface [2], [12], [13].

Formulating the problem of image segmentation as an energy minimization (or a posterior distribution maximization) to be solved by graph cut can be justified by the theory of Markov random field (MRF) [15], [16]. A lot of successful results have been proposed in recent years such as the interactive object extraction [17], [18], [19] and the iterative Grab-cut system [3], where an efficient min-cut/max-flow algorithm proposed in [20] is adopted to minimize the energy function. This min-cut/max-flow algorithm is guaranteed to find the global optimal for certain types of energy functions which satisfy the property that they are functions of binary variables, submodular, and can be written as the sum of terms involving at most three variables at a time [21]. For energy functions with multi-label variables, approximate solution can be obtained by using the algorithm proposed in [22] which utilize a sequence of binary moves such as alpha-expansion, alpha-beta swap and k-jumps, etc.. Although there are efficient polynomial time min-cut/max-flow algorithms [20], the types of energy functions it can minimize are still limited [21]. A more general but less efficient algorithm, which can sample from arbitrary posterior distributions and thus can minimize a more general set of energy functions, is the Swendsen-Wang cut [23], [24] and the generalized m-way Swendsen-Wang cut [25].

Both the variational energy minimization approach and the graph energy minimization approach share the same methodology: formulating an energy function and solving the resulting optimization problem. What make them different are the different optimization strategies being used. The variational energy minimization can be converted to a PDE and solved by FDM [4], [1], FEM [6] and level set [2], while the graph energy minimization could be solved by min-cut/max-flow algorithms such as the one in [20] and the Swendsen-Wang cut [23], [24], [25]. What kind of optimization scheme is more suited is usually determined by the type of objective function. The objective function is also a main factor determining the quality of the segmentation

results. Therefore, it is misleading to only ask question "*which method, graph cut or level set, produces better image segmentation results*", since it all depends on the objective function. While it is extremely important to study various optimization schemes, this paper mainly focuses on a better and justifiable energy function formulation.

We propose a novel local-global variational energy functional for the problem of extracting the foreground OOI from static images. The novelty comes from the incorporation of a global image data likelihood potential that seeks for a global description of all the pixels in the image. This addresses the problem that during the iterations the GMM model for each region (e.g. foreground or background) is estimated locally from the pixels in the currently estimated region which is in general different from the true region. Basically on one hand the estimated region may contain only a portion of the pixels that belong to the true region, and on the other hand it may contain pixels that do not belong to the true region. Note that the proposed variational energy functional can not be optimized by a graph-cut technique because it is not clear how to incorporate the curve energy term into a graph-cut optimization scheme. We choose to use a level set approach and a novel quasi-semi-supervised EM algorithm to carry out the optimization.

III. Robust variational formulation

The definition of "homogeneity" is critical for any image segmentation algorithm. It is natural to model the homogeneity of an image region using a probabilistic distribution. For example, a Gaussian distribution on the pixel intensity was adopted in [1], and a learned Gaussian mixture model for each texture region was adopted in [2].

Our goal is to extract a foreground object from the background. It is not realistic to assume the foreground object or the background region is a single Gaussian distribution. We instead model the feature distributions of both the foreground and the background regions as Gaussian mixtures. Denote the foreground image as $\mathcal{F}$, the background image as $\mathcal{B}$, the image data $\mathcal{I} = \mathcal{F} \cup \mathcal{B}$ and $\mathbf{u}(x, y)$ as the feature vector at image coordinate (x, y), we have

$$\begin{aligned} P_{\mathcal{F}}(\mathbf{u}(x,y)) &= P(\mathbf{u}(x,y)|(x,y) \in \mathcal{F}) = \sum_{i=1}^{K_{\mathcal{F}}} \pi_i^{\mathcal{F}} \mathcal{N}(\mathbf{u}(x,y)|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}}) \\ P_{\mathcal{B}}(\mathbf{u}(x,y)) &= P(\mathbf{u}(x,y)|(x,y) \in \mathcal{B}) = \sum_{i=1}^{K_{\mathcal{B}}} \pi_i^{\mathcal{B}} \mathcal{N}(\mathbf{u}(x,y)|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}}), \end{aligned} \tag{1}$$

where π_i, $\vec{\mu}_i$ and Σ_i are, respectively, the mixture weight, the mean and the covariance of the corresponding Gaussian components, and $K_{\mathcal{F}}$ and $K_{\mathcal{B}}$ represent the number of Gaussian components in each of the Gaussian mixtures.

Assuming the image pixels are drawn *i.i.d.* from the two Gaussian mixtures, the image data likelihood is simply a mixture model of the foreground and background distributions, that is,

$$P_{\mathcal{I}}(\mathbf{u}(x,y)) = \omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}(x,y)) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u}(x,y)),\ \ s.t.,\ \ \omega_{\mathcal{F}} + \omega_{\mathcal{B}} = 1, \tag{2}$$

where $\omega_{\mathcal{F}} = P((x,y) \in \mathcal{F})$ and $\omega_{\mathcal{B}} = P((x,y) \in \mathcal{B})$ are the prior probabilities that a pixel is drawn from the foreground and background, respectively.

A. Local region potential

Denote $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ as the estimated foreground and background regions respectively. Let $\mathcal{I}$ denote the whole image data, that is, $\mathcal{I} = \mathcal{A}_{\mathcal{F}} \cup \mathcal{A}_{\mathcal{B}}$. The quality of the estimation can be evaluated by the joint likelihood probabilities of he foreground and background pixels, 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 of Equation 3, we obtain the local region likelihood potential as

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

Our local region potential energy in Equation 4 is more general than the energy function adopted in [1], [2] since we have incorporated the prior probabilities of the foreground and background. In the case where we have no prior knowledge about $\omega_{\mathcal{F}}$ and $\omega_{\mathcal{B}}$ and set them both to $\frac{1}{2}$, the local region potential in Equation 4 boils down to what is used in [1], [2].

B. Global image data likelihood potential

The maximization of $\mathbf{E}_h$ with respect to the regions $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ and the probability distribution is a *chicken-and-egg* problem. If we know $\omega_{\mathcal{F}}$, $\omega_{\mathcal{B}}$, $P_{\mathcal{F}}$ and $P_{\mathcal{B}}$, we can easily identify the optimal $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$, and vice versa. In practice, the regions and the probability parameters are solved alternatively. At each iteration, $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ are fixed first while the probability parameters are

solved to maximize $\mathbf{E}_h$. Then the probability parameters are fixed while the regions become unknowns to solve for.

Notice that Equation 4 only independently evaluates the fitness of the estimated foreground and background region. When the estimated foreground and background regions are close to the ground truth, Equation 4 gives the maximum likelihood estimation for the probability model which makes perfect sense. But in practice, $\mathcal{A}_\mathcal{F}$ and $\mathcal{A}_\mathcal{B}$ are usually quite different from the ground truth during the iteration process. In other words, $\mathcal{A}_\mathcal{F}$ may not contain all the pixels in the foreground, and furthermore, it may contain pixels which belong to the background. The same problem exists with $\mathcal{A}_\mathcal{B}$. This affects the accuracy of the probability model parameters which in turn affects the subsequent segmentation. To address this problem, we propose to incorporate a global image data likelihood that seeks for a global description of the entire image data.

Since the image pixels can be regarded as *i.i.d.* samples drawn from $P_\mathcal{I}(\mathbf{u}(x,y))$, the global image data likelihood is the following:

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

By taking the logarithm, the global image data likelihood potential is finally obtained as

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

C. Boundary potential

Image edges provide strong cues for segmentation. There has been a significant literature which incorporate edge information into a variational energy function such as the SNAKES [4], the active contour model with balloons [6] and the geodesic active contour [26], to list a few.

Since the geodesic active contour overcomes some of the intrinsic limitations of SNAKES, we adopt a similar formulation to obtain the optimal boundary $\Gamma(c) : c \in [0,1] \rightarrow (x,y) \in \mathbf{R}^2$, which is a closed curve between the region $\mathcal{A}_\mathcal{F}$ and the region $\mathcal{A}_\mathcal{B}$ such that $\Gamma(c) = \mathcal{A}_\mathcal{F} \cap \mathcal{A}_\mathcal{B}$. The energy term corresponding to edge information is

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

where $\mathbf{g}_x$ and $\mathbf{g}_y$ are the image gradient at the image coordinate (x,y), and $\dot{\Gamma}(c)$ is the first order derivative of the boundary curve. Minimizing $\mathbf{E}_e(\Gamma(c))$ will align the boundary curve $\Gamma(c)$ to the image pixels with the maximum image gradient while $\dot{\Gamma}(c)$ will impose the first order smoothness constraint on the boundary curve.

D. Boundary, region and data likelihood synergism

There has been a lot of work that formulate the image segmentation as a variational energy minimization problem such as the region competition [1] and the geodesic active region [2]. We also use a variational energy minimization approach. Our energy functional is

$$\begin{aligned}
\mathbf{E}_p(\Gamma(c), P_{\mathcal{I}}) &= \alpha\mathbf{E}_e - \beta\mathbf{E}_h - \gamma\mathbf{E}_l \\
&= \alpha \underbrace{\int_0^1 \frac{1}{1+|\mathbf{g}_x(\Gamma(c))|+|\mathbf{g}_y(\Gamma(c))|}|\dot{\Gamma}(c)|dc}_{\mathbf{E}_e} \\
&- \beta \underbrace{\left(\int_{\mathcal{A}_{\mathcal{F}}} \{\log P_{\mathcal{F}}(\mathbf{u}) + \log\omega_{\mathcal{F}}\} + \int_{\mathcal{A}_{\mathcal{B}}} \{\log P_{\mathcal{B}}(\mathbf{u}) + \log\omega_{\mathcal{B}}\}\right)}_{\mathbf{E}_h} \\
&- \gamma \underbrace{\int_{\mathcal{A}_{\mathcal{F}}\cup\mathcal{A}_{\mathcal{B}}} \log\{\omega_{\mathcal{F}}P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}}P_{\mathcal{B}}(\mathbf{u})\}}_{\mathbf{E}_l}, \qquad (8)
\end{aligned}$$

where α, β and γ are positive numbers with $\alpha+\beta+\gamma = 1$. They are intended to balance different energy terms.

Compared with the formulation of the potential energy in the region competition [1] and the geodesic active region [2], the uniqueness of our formulation is that none of their formulation incorporated the *global image data likelihood potential*. Moreover, the region competition approach [1] assumes a Gaussian distribution for each homogenous region while we use Gaussian mixtures to model the foreground and background. In the geodesic active region approach [2], the foreground and background distributions are pre-trained which renders the potential energy to be only dependent on the boundary curve $\Gamma(c)$.

E. General machine learning problem behind the joint local-global energy

The estimation of the image data model by minimizing the joint global likelihood energy and the local region energy indeed reveals a very interesting machine learning problem, i.e., learning with inaccurately labeled data set. It can be stated more rigorously as the following proposition.

Proposition 3.1: Let $\mathcal{D} = \{d_i|1 = 1\ldots n\}$ be a ***i.i.d.*** data set drawn from a mixture data model $P(\mathbf{d}|\mathbf{\Theta}) = \omega_1 P_1(\mathbf{d}|\mathbf{\Theta}_1) + \omega_2 P_2(\mathbf{d}|\mathbf{\Theta}_2)$, where $\omega_1+\omega_2 = 1$ and $\mathbf{\Theta} = \{\omega_i, \mathbf{\Theta}_i, i = 1, 2\}$. Assume $\mathcal{L} = \{l_i|l_i \in \{1, 2\}, i = 1\ldots n\}$ be the unknown ground truth binary label set indicating that each d_i is drawn from $P_{l_i}(\mathbf{d}|\mathbf{\Theta}_{l_i})$. Suppose we have an inaccurate label set $\mathcal{Z} = \{z_i|z_i \in$

$\{1,2\}, i = 1 \ldots n\}$ where an unknown portion $\mathcal{E} = \{z_i | z_i \neq l_i\}$ are incorrectly labeled. Then the problem is that given $\mathcal{D}$ and $\mathcal{Z}$, how could we robustly estimate $P_{\mathcal{D}}(\mathbf{d}|\mathbf{\Theta})$, or more concrete the model parameters $\mathbf{\Theta}$?

In principle, this is a parameter estimation problem in between of purely supervised parameter learning and purely unsupervised parameter learning, since we do have labeled data set but the labels are not accurate. Just considering the situation that all the labels are correct, i.e., $\mathcal{E} = \emptyset$, we can easily estimate $\mathbf{\Theta}$ by the routine maximum likelihood estimation (MLE). Without loss of any generality, if over 50% of the data points have been erroneously labeled, the labels will not provide any helpful information for the estimation of the parameters since a random guess of the label may do better than the labels provided. In this case, the parameters of the data model can only be estimated by unsupervised learning algorithm such as the popular EM algorithm [27].

Denote $\mathcal{D}_1 = \{d_i | z_i = 1\}$ and $\mathcal{D}_2 = \{d_i | z_i = 2\}$, we have $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$. Mathematically, the purely supervised learning targets on maximizing the following log likelihood function in Equation 9.

$$\log P_{\mathcal{D}_1 \mathcal{D}_2} = \sum_{\mathcal{D}_1} \{\log \omega_1 + \log P_1(\mathbf{d}|\mathbf{\Theta}_1)\} + \sum_{\mathcal{D}_2} \{\log \omega_2 + \log P_2(\mathbf{d}|\mathbf{\Theta}_2)\} . \tag{9}$$

It is easy to figure out that Equation 9 exactly corresponds to the local region potential in our variational energy forumlation in Equation 8 for the image segmentation problem. On the other hand, purely unsupervised learning aims at maximizing the following joint data log likelihood function in Equation 10.

$$\log P_{\mathcal{D}} = \sum_{\mathcal{D}} \{\omega_1 P_1(\mathbf{d}|\mathbf{\Theta}_1) + \omega_2 P_2(\mathbf{d}|\mathbf{\Theta}_2)\} . \tag{10}$$

It is also easy to figure out that Equation 10 exactly corresponds to the global data likelihood potential in Equation 8. For the problem stated in the proposition 3.1, if $\mathcal{E} \neq \emptyset$ and there is a significant part (e.g., over 60%) of the labels which have been correctly labeled, both the purely supervised learning scheme and purely unsupervised learning scheme are not suitable. Intuitively, purely supervised learning scheme may result in a very biased estimation due to the erroneously labeled data points. On the other hand, purely unsupervised learning scheme totally ignores the useful information from the correctly labeled data points. Ideally, we should effectively utilize the correctly labeled data and reduce the effects of the erroneously labeled data to the minimum for the robust estimation of the model parameters. To achieve this, we propose to maximize the

following combined log likelihood function, i.e.,

$$
\begin{aligned}
F_{\mathcal{D}} &= \alpha \log P_{\mathcal{D}_1\mathcal{D}_2} + (1-\alpha)\log P_{\mathcal{D}} \qquad (11)\\
&= \alpha \left\{ \sum_{\mathcal{D}_1} \{\log \omega_1 + \log P_1(\mathbf{d}|\mathbf{\Theta}_1)\} + \sum_{\mathcal{D}_2} \{\log \omega_2 + \log P_2(\mathbf{d}|\mathbf{\Theta}_2)\} \right\} \\
&+ (1-\alpha) \left\{ \sum_{\mathcal{D}} \{\omega_1 P_1(\mathbf{d}|\mathbf{\Theta}_1) + \omega_2 P_2(\mathbf{d}|\mathbf{\Theta}_2)\} \right\}, \qquad (12)
\end{aligned}
$$

where $0 \leq \alpha \leq 1$ should be set to make a balancing between the supervised learning and unsupervised learning scheme based on our confidence about the correctness of the labels. It is intuitive to see that maximize this combined log likelihood function will fit the data model locally with the labeled data and at the same time seek for a global description of the whole data to reduce the effects of those erroneously labeled data to the minimum. We name this type of problem as a *quasi-semi-supervised* learning problem.

IV. Energy minimization algorithms

Since we do not have a pre-specified image data model $P_{\mathcal{I}}(\mathbf{u})$, it is obvious that the variational energy functional we formulated in Equation 8 relies on two sets of functions, i.e., the boundary curve $\Gamma(c)$, and the the image data model $P_{\mathcal{I}}(\mathbf{u})$. Therefore, we propose a two step iterative process to minimize the energy functional, i.e., at one step, with fixed $P_{\mathcal{I}}(\mathbf{u})$, we minimize the energy with respect to the $\Gamma(c)$. While at the other step, we minimize the energy functional with respect to $P_{\mathcal{I}}(\mathbf{u})$ with a fixed boundary $\Gamma(c)$. Each step will guarantee to minimize the variational energy, we present more details of the two steps as follows.

A. Boundary optimization by level set

In the first step of our iterative minimization scheme, we fix $P_{\mathcal{F}}(\mathbf{u})$, $P_{\mathcal{B}}(\mathbf{u})$, $\omega_{\mathcal{F}}$ and $\omega_{\mathcal{B}}$, and minimize the functional with respect to $\Gamma(c)$. This can be achieved by gradient decent, e.g., take the variation of $\mathbf{E}_p(\Gamma(c), P_{\mathcal{F}}, P_{\mathcal{B}})$ with respect to $\Gamma(c)$, we have

$$
\begin{aligned}
\frac{\partial \mathbf{E}_p}{\partial \Gamma(c)} &= \beta \log \left[\frac{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}(\Gamma(c)))}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u}(\Gamma(c)))} \right] \cdot \vec{n}(\Gamma(c)) \\
&+ \alpha \left[G(\Gamma(c)) \mathcal{K}(\Gamma(c)) - \nabla G(\Gamma(c)) \cdot \vec{n}(\Gamma(c)) \right] \cdot \vec{n}(\Gamma(c)), \qquad (13)
\end{aligned}
$$

where $\vec{n}(\cdot)$ represents the normal line pointing outwards from the boundary curve $\Gamma(c)$, $\mathcal{K}(\cdot)$ is the curvature, and all the function values should be evaluated on the boundary curve $\Gamma(c)$. One

interesting observation here is that the form of the partial variation in Equation 13 is almost the same as that in [2] except the mixture weights $\omega_{\mathcal{F}}$ and $\omega_{\mathcal{B}}$. This means that the image data likelihood potential $\mathbf{E}_l$ does not affect the partial variation of the energy functional with respect to the curve. This is easy to understand because the $\mathbf{E}_l$ is evaluated on the whole image, it does not rely on the boundary curve $\Gamma(c)$.

We propose to use level set to implement the above partial derivative equations, i.e., at each time instant t during the optimization of the curve, $\Gamma(c,t)$ is represented as the zero level set of a 2 dimensional function or surface $\varphi(x,y,t)$, i.e., $\Gamma(c,t) := \{(x,y)|\varphi(x,y,t)=0\}$. Following the literature [2], [26], we define $\varphi(x,y,t)$ to be a signed distance function, i.e.,

$$\varphi(x,y,t) = \begin{cases} d((x,y),\Gamma(c,t)) & ,(x,y) \in \mathcal{A}_{\mathcal{F}} \setminus \Gamma(c,t) \\ 0 & ,(x,y) \in \Gamma(c,t) \\ -d((x,y),\Gamma(c,t)) & ,(x,y) \in \mathcal{A}_{\mathcal{B}} \setminus \Gamma(c,t) \end{cases} \tag{14}$$

where $d(\cdot)$ is the Euclidean distance from the point (x,y) to $\Gamma(c,t)$ which is defined as the shortest possible distance from (x,y) to any points in $\Gamma(c,t)$. We have

$$\begin{aligned} \frac{\partial\varphi(x,y,t)}{\partial t} &= \beta \log\left[\frac{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}(x,y))}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u}(x,y))}\right] |\nabla\varphi(\cdot)| \\ &+ \alpha\left[G(x,y)\mathcal{K}(x,y) - \nabla G(x,y)\cdot\frac{\nabla\varphi(\cdot)}{|\nabla\varphi(\cdot)|}\right]|\nabla\varphi(\cdot)| \end{aligned} \tag{15}$$

where

$$\mathcal{K}(x,y) = \frac{\varphi_{xx}\varphi_y^2 - 2\varphi_{xy}\varphi_x\varphi_y + \varphi_{yy}\varphi_x^2}{(\varphi_x^2+\varphi_y^2)^{\frac{3}{2}}}, \tag{16}$$

among which φ_x and φ_y, and φ_{xx}, φ_{yy} and φ_{xy} are the set of first order partial derivatives and the set of second order partial derivatives of $\varphi(x,y,t)$, respectively.

The evolution of $\varphi(x,y,t)$ over time t is then implemented by replacing the derivatives by discrete differences, i.e., the partial derivative with respect to t is approximated by forward differences and the partial derivative with respect to x and y are approximated by central differences. In principle, the evolution of the surface is evaluated by

$$\begin{aligned} \varphi(x,y,t+\tau) &= \varphi(x,y,t) + \tau\cdot\left\{\beta\log\left[\frac{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}(x,y))}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u}(x,y))}\right]|\nabla\varphi(\cdot)|\right. \\ &+ \left.\alpha\left[G(x,y)\mathcal{K}(x,y) - \nabla G(x,y)\cdot\frac{\nabla\varphi(\cdot)}{|\nabla\varphi(\cdot)|}\right]|\nabla\varphi(\cdot)|\right\}, \end{aligned} \tag{17}$$

where τ is the discrete time step. We have $\Gamma(c,t+\tau) = \{(x,y)|\varphi(x,y,t+\tau)=0\}$.

B. Image data model estimation

In the second step of our iterative minimization scheme, we fix the the boundary curve $\Gamma(c)$ and minimize the energy functional with respect to $P_{\mathcal{F}}(\mathbf{u})$, $P_{\mathcal{B}}(\mathbf{u})$, $\omega_{\mathcal{F}}$ and $\omega_{\mathcal{B}}$ at the same time. In other words, by fixing $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$, we minimize the functional with respect to $P_{\mathcal{I}}(\mathbf{u})$. In principle, this involves to minimize the variational energy with respect to all the parameters $\mathbf{\Theta}$ of $P_{\mathcal{I}}(\mathbf{u})$, i.e.,

$$\mathbf{\Theta} = \left\{\omega_{\mathcal{F}}, \omega_{\mathcal{B}}, \{\pi_i^{\mathcal{F}}, \vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}}\}_{i=1}^{K_{\mathcal{F}}}, \{\pi_i^{\mathcal{B}}, \vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}}\}_{i=1}^{K_{\mathcal{B}}}\right\}. \tag{18}$$

Take the derivative of $\mathbf{E}_p$ with respect to each parameter in $\mathbf{\Theta}$, we have

$$\frac{\partial \mathbf{E}_p}{\partial \omega_{\mathcal{F}}} = \beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{1}{\omega_{\mathcal{F}}} + \gamma \int_{\mathcal{I}} \frac{P_{\mathcal{F}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} \tag{19}$$

$$\frac{\partial \mathbf{E}_p}{\partial \omega_{\mathcal{B}}} = \beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{1}{\omega_{\mathcal{B}}} + \gamma \int_{\mathcal{I}} \frac{P_{\mathcal{B}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} \tag{20}$$

$$\frac{\partial \mathbf{E}_p}{\partial \pi_i^{\mathcal{F}}} = \beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{\omega_{\mathcal{F}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\omega_{\mathcal{F}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}} \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} \tag{21}$$

$$\begin{aligned}\frac{\partial \mathbf{E}_p}{\partial \mu_i^{\mathcal{F}}} &= \beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{\omega_{\mathcal{F}} \pi_i^{\mathcal{F}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})(\Sigma_i^{\mathcal{F}})^{-1}(\mathbf{u} - \vec{\mu}_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} \\ &+ \gamma \int_{\mathcal{I}} \frac{\omega_{\mathcal{F}} \pi_i^{\mathcal{F}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}} \Sigma_i^{\mathcal{F}})(\Sigma_i^{\mathcal{F}})^{-1}(\mathbf{u} - \vec{\mu}_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}\end{aligned} \tag{22}$$

$$\begin{aligned}\frac{\partial \mathbf{E}_p}{\partial \Sigma_i^{\mathcal{F}}} &= \beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{\omega_{\mathcal{F}} \pi_i^{\mathcal{F}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})(\Sigma_i^{\mathcal{F}})^{-1} \left[(\mathbf{u} - \vec{\mu}_i^{\mathcal{F}})(\mathbf{u} - \vec{\mu}_i^{\mathcal{F}})^T (\Sigma_i^{\mathcal{F}})^{-1} - \mathbf{I}\right]}{2\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} \\ &+ \gamma \int_{\mathcal{I}} \frac{\omega_{\mathcal{F}} \pi_i^{\mathcal{F}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})(\Sigma_i^{\mathcal{F}})^{-1} \left[(\mathbf{u} - \vec{\mu}_i^{\mathcal{F}})(\mathbf{u} - \vec{\mu}_i^{\mathcal{F}})^T (\Sigma_i^{\mathcal{F}})^{-1} - \mathbf{I}\right]}{2[\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})]}\end{aligned} \tag{23}$$

$$\frac{\partial \mathbf{E}_p}{\partial \pi_i^{\mathcal{B}}} = \beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{\omega_{\mathcal{B}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\omega_{\mathcal{B}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}} \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} \tag{24}$$

$$\begin{aligned}\frac{\partial \mathbf{E}_p}{\partial \mu_i^{\mathcal{B}}} &= \beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{\omega_{\mathcal{B}} \pi_i^{\mathcal{B}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})(\Sigma_i^{\mathcal{B}})^{-1}(\mathbf{u} - \vec{\mu})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} \\ &+ \gamma \int_{\mathcal{A}_{\mathcal{B}} \cup \mathcal{A}_{\mathcal{B}}} \frac{\omega_{\mathcal{B}} \pi_i^{\mathcal{B}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}} \Sigma_i^{\mathcal{B}})(\Sigma_i^{\mathcal{B}})^{-1}(\mathbf{u} - \vec{\mu})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}\end{aligned} \tag{25}$$

$$\begin{aligned}\frac{\partial \mathbf{E}_p}{\partial \Sigma_i^{\mathcal{B}}} &= \beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{\omega_{\mathcal{B}} \pi_i^{\mathcal{B}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})(\Sigma_i^{\mathcal{B}})^{-1} \left[(\mathbf{u} - \vec{\mu})(\mathbf{u} - \vec{\mu})^T (\Sigma_i^{\mathcal{B}})^{-1} - \mathbf{I}\right]}{2\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} \\ &+ \gamma \int_{\mathcal{I}} \frac{\omega_{\mathcal{B}} \pi_i^{\mathcal{B}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})(\Sigma_i^{\mathcal{B}})^{-1} \left[(\mathbf{u} - \vec{\mu})(\mathbf{u} - \vec{\mu})^T (\Sigma_i^{\mathcal{B}})^{-1} - \mathbf{I}\right]}{2[\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})]},\end{aligned} \tag{26}$$

where $\mathbf{I}$ is the identity matrix. Set all the derivatives to zero and after some mathematical

manipulation, we easily come up with the following fixed-point equations, i.e.,

$$\omega_{\mathcal{F}}^{*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{F}}} 1 + \gamma \int_{\mathcal{I}} \frac{2\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\gamma \int_{\mathcal{I}} \frac{P_{\mathcal{F}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}} \tag{27}$$

$$\omega_{\mathcal{B}}^{*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{B}}} 1 + \gamma \int_{\mathcal{I}} \frac{2\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\gamma \int_{\mathcal{I}} \frac{P_{\mathcal{B}}(\mathbf{u})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}} \tag{28}$$

$$\pi_i^{\mathcal{F}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{\pi_i^{\mathcal{F}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{2\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}} \tag{29}$$

$$\vec{\mu}_i^{\mathcal{F}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{\mathbf{u}\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathbf{u}\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}} \tag{30}$$

$$\Sigma_i^{\mathcal{F}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{(\mathbf{u}-\vec{\mu}_i^{\mathcal{F}})(\mathbf{u}-\vec{\mu}_i^{\mathcal{F}})^T \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{(\mathbf{u}-\vec{\mu}_i^{\mathcal{F}})(\mathbf{u}-\vec{\mu}_i^{\mathcal{F}})^T \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{F}}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{F}}, \Sigma_i^{\mathcal{F}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}} \tag{31}$$

$$\pi_i^{\mathcal{B}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{\pi_i^{\mathcal{B}} \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{2\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}} \tag{32}$$

$$\vec{\mu}_i^{\mathcal{B}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{\mathbf{u}\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathbf{u}\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}} \tag{33}$$

$$\Sigma_i^{\mathcal{B}*} = \frac{\beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{(\mathbf{u}-\vec{\mu}_i^{\mathcal{B}})(\mathbf{u}-\vec{\mu}_i^{\mathcal{B}})^T \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{(\mathbf{u}-\vec{\mu}_i^{\mathcal{B}})(\mathbf{u}-\vec{\mu}_i^{\mathcal{B}})^T \mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}{\beta \int_{\mathcal{A}_{\mathcal{B}}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})} + \gamma \int_{\mathcal{I}} \frac{\mathcal{N}(\mathbf{u}|\vec{\mu}_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}})}{\omega_{\mathcal{F}} P_{\mathcal{F}}(\mathbf{u}) + \omega_{\mathcal{B}} P_{\mathcal{B}}(\mathbf{u})}}, \tag{34}$$

which are also subject to the constraints that

$$\omega_{\mathcal{F}}^{*} + \omega_{\mathcal{B}}^{*} = 1, \quad \sum_{i=1}^{K_{\mathcal{F}}} \pi_i^{\mathcal{F}} = 1, \quad \sum_{i=1}^{K_{\mathcal{B}}} \pi_i^{\mathcal{B}} = 1. \tag{35}$$

Therefore, we must ensure that we normalize these weights at each iteration of the fixed-point iterations.

This set of fixed-point equations can be interpreted as a robust quasi-semi-supervised EM algorithm for Gaussian mixture models, where we have inaccurate labels of the data in a 2-class classification problem, and each class can be represented by a Gaussian mixture model. It turns out that the robust estimation of the data distribution, and thus the probabilistic distribution for each of the class could be achieved by fixed-point iteration similar to that in Equation 27 to Equation 34. This is just a specific result on Gaussian mixture models on the general machine learning problem we have discussed in Section III-E.

Here the foreground and background image pixels are the two classes we would want to discriminate, and $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ can be regarded as the inaccurate labeling of the foreground and background pixels. The fixed-point equations we have derived try to make a balancing between the estimation from the labeled data and the unsupervised estimation. The erroneous labeled data will be given less weight during the fixed-point iteration. This can be easily observed in Equation 30, the first integral of the numerator over $\mathcal{A}_{\mathcal{F}}$ is in fact the estimation from the inaccurately labeled data, and the second integration of the numerator over $\mathcal{I} = \mathcal{A}_{\mathcal{F}} \cup \mathcal{A}_{\mathcal{B}}$ is a soft classification of the image pixels by the current estimation of the data likelihood model. Those image pixels which have been labeled to be in $\mathcal{A}_{\mathcal{F}}$, and which have also been classified with high confidence as foreground pixels will be given more weight. This will result in a more robust estimation of the data distribution since the effects of those erroneously labeled data will be suppressed. This has also been demonstrated in our experiments in Section V.

V. Experiments: extracting objects at the focus of attention

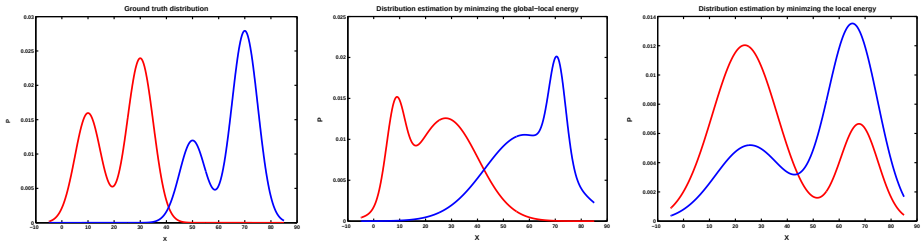
Although the formulation of the proposed method is very general to handle image segmentation in a general setting, we focus on a more specific application, i.e., the extraction of object at the focus of attention. The basic assumption is that when one takes an image of an object of interest, he will usually locate it in the center of the image. This weak assumption does not affect our formulation, but makes the fully automatic extraction possible.

A. Validation of minimizing the local-global energy on numerical example

We use a synthetic numerical example to demonstrate the effectiveness of minimizing the local-global energy in the problem of model estimation with inaccurately labeled data. The ground truth data model is

$$\begin{aligned} P(\mathbf{d}|\mathbf{\Theta}) &= \omega_1 P_1(\mathbf{d}|\mathbf{\Theta}_1) + \omega_2 P_2(\mathbf{d}|\mathbf{\Theta}_2) \\ &= 0.5\,(0.4\mathcal{N}(\mathbf{d}|10,25) + 0.6\mathcal{N}(\mathbf{d}|30,25)) \\ &+ 0.5\,(0.3\mathcal{N}(\mathbf{d}|50,25) + 0.7\mathcal{N}(\mathbf{d}|70,25)) \end{aligned} \tag{36}$$

where $\mathcal{N}(\mathbf{d}|\mu,\sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2. Then $\mathbf{\Theta}$ represents the set of parameters for all the Gaussian mixture models. We then randomly draw a set $\mathcal{D}$ of 20000 data samples from the data model and during the sampling process, we also



(a) Ground truth (b) Local-global energy (c)Local energy

Fig. 1. Comparison of minimizing the local energy and joint local-global energy function. From left to right are the estimation results of the ground truth, the result of minimizing the local-global energy and the result of minimizing only the local energy.

recorded the set $\mathcal{L}$ of ground-truth labels, which indicates whether a sample is from P_1 or P_2, and we denote $\mathcal{L}_1 = \{l_i = 1\}$ and $\mathcal{L}_2 = \{l_i = 2\}$. To simulate the situation of erroneous labeling, we randomly exchange 30% of the labels between $\mathcal{L}_1$ and $\mathcal{L}_2$. We then denote the exchanged label set as $\mathcal{Z}_1$ and $\mathcal{Z}_2$ which are supposed to be the known condition.

We then compared the model estimation results of only maximizing the local log likelihood (or equivalently minimizing the local energy due to the negative sign) in Equation 9 and the results of maximizing the local-global log likelihood (or equivalently minimizing the local-global energy due to the negative sign) in Equation 10. Figure 1(b) shows the model estimated by minimizing the local-global energy function, Figure 1(c) shows the model estimated by only minimizing the local energy function. Compared with the ground truth model presented in Figure 1(a), we can easily see that the model estimated by local-global energy minimization is far more close to the real model than the model estimated by the local energy minimization.

Since 30% erroneous labels is significant, we set α to be 0.05 to balance more toward the global energy function. In the experiments, the local-global energy minimization is performed by fixed point iteration similar to what has been derived in Section IV-B, namely quasi-semi-supervised EM. The local energy minimization is performed by applying the classical EM algorithm [27] to fitting the two Gaussian mixtures independently on the two data sets induced by $\mathcal{Z}_1$ and $\mathcal{Z}_2$. For both algorithms, we randomly choose 10 different initializations and the best results on the 10 runs are adopted for comparison. We have extensively run the algorithm with different setting of the ground truth data model parameters. We generally observe similar results as shown in Figure 1.

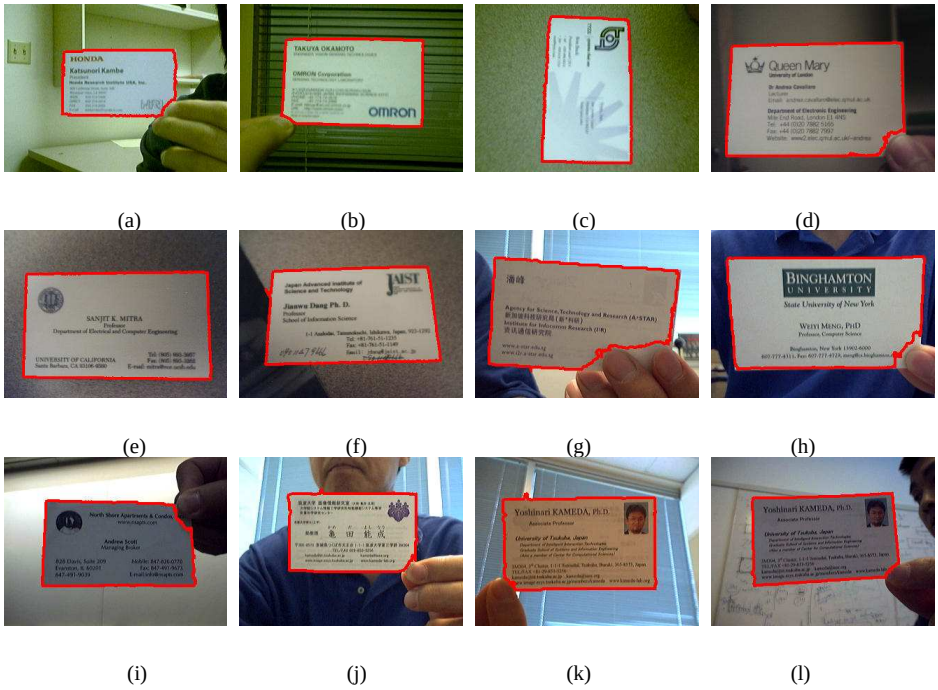


Fig. 2. Card segmentation results.

B. *Automatic extraction of business card from images*

We have constructed a fully automatic real-time system to extract business card from still images of unconstrained background. We can then rectify the shape of the extracted business card to be a rectangle with the correct physical aspect ratio by using techniques similar to that in [28]. We further enhance the rectified image, e.g., enhance the contrast of the rectified image by transforming it through a "S" shaped Hermite curve interpolated according the intensity distribution of the image pixels.

In summary, the whole system is composed of three sub-systems, namely the segmentation subsystem, the shape rectification subsystem and the image enhancement subsystem.

1) Business card segmentation: The segmentation of the business card in the image is achieved by the proposed algorithm. The output of the sub-system is a clock-wise chain code of the image coordinates of the closed boundary of the business card region identified, along with the labeling of whether a pixel belongs to the business card or background. Some implementation details and explanation are as follows:

- INPUT: The input is a color image, and the image feature vector $\mathbf{u}$ adopted is a five dimensional vector $\{L, U, V, x, y\}$, where L, U and V are the color pixel values in the LUV color space and x and y are the coordinates of the pixels in the image. We adopt the LUV color space because it was specially designed to best approximate perceptually uniform color spaces [29]. That will facilitate to obtain a meaningful segmentation as the perceived color difference in the LUV space is very coherent to be an Euclidean metric.
- MODEL: The foreground object model $P_{\mathcal{F}}$ is a 2-component mixture of Gaussian, which models the bright sheet and dark characters of most of the business card. The background model $P_{\mathcal{B}}$ is a 8-component mixture of Gaussian, which should cover most of the pixels located in the boundary of the image coordinate.
- INITIALIZATION OF SURFACE: The initial level set surface is initialized by a signed distance transform with respect to a rectangle located in the center of the image with length and width of $\frac{1}{8}$ of the image width and length.
- INITIALIZATION OF FOREGROUND MODEL: Firstly, we sort the pixels inside the initial rectangle according to their intensity value L. Then we take $K_{\mathcal{F}} = 2$ average values of the 5-dimensional feature vectors of the lightest 10% pixels and the darkest 10% pixels respectively as the seeds for the mean-shift mode seeking on the feature space of the whole image. The two modes obtained are then adopted as the initialization of $\vec{\mu}_1^{\mathcal{F}}$ and $\vec{\mu}_2^{\mathcal{F}}$. The mixture weights $\pi_1^{\mathcal{F}}$ and $\pi_2^{\mathcal{F}}$ are both initialized to be 0.5. Each covariance matrix $\Sigma_i^{\mathcal{F}}$ is initialized as the same diagonal covariance matrix, i.e., the variances of the spatial components (x, y) are initialized as $\frac{1}{5}$ of the image width and height, respectively. The variances of the color components $\{L, U, V\}$ are all initialized as 25. Note we do not make much effort to tune these initialization parameters.
- INITIALIZATION OF BACKGROUND MODEL: The $K_{\mathcal{B}} = 8$ average feature vectors of pixels inside eight 10×10 rectangles, which are circled around the margin of the image, are adopted as the initialization of the mean-shift mode seeking algorithm in the full image feature space. The eight recovered feature modes are then adopted as the initialization of each $\vec{\mu}_i^{\mathcal{B}}$ of $P_{\mathcal{B}}(\mathbf{u})$. The covariance matrices $\Sigma_i^{\mathcal{B}}$, $i = 1, \ldots, 8$ have the same initialization with those of the foreground model $P_{\mathcal{F}}(\mathbf{u})$. All the $\pi_i^{\mathcal{B}}$s are set to be $\frac{1}{8}$.

(a) (b) (c)

Fig. 3. Some failure cases of the variational energy formulation without the global image data likelihood energy.

- INITIALIZATION OF FOREGROUND/BACKGROUND MIXTURE WEIGHT: The mixture weights $\omega_{\mathcal{F}}$ and $\omega_{\mathcal{B}}$ are initialized to be equal to $\frac{1}{2}$.
- CONVERGENCE CRITERION: Whenever the foreground region has less than 1% change in two consecutive iterations, we consider the algorithm is converged. However, this criterion is very rough but it meets the requirement of the business card extraction system. We also set a maximum iteration number of 30 in case the 1% change criterion can not be achieved within the processing time we can tolerate.

Note that these settings are applied to all the experiments performed and reported in this paper. We present some segmentation results of different business card in various background in Figure 2. The closed boundary of the business card is overlayed in red. We can generally obtain satisfactory segmentation results, i.e, we achieve over 95% successful rate on over 300 images tested. We regard a segmentation result to be successful if it is almost matched with the region which would be segmented by human perception.

For comparison, we have also run the algorithm from the variational energy formulation without the global image data likelihood included. Now in the step of estimating the distribution $P_{\mathcal{F}}$ and $P_{\mathcal{B}}$ ($\omega_{\mathcal{F}}$ and $\omega_{\mathcal{B}}$ are not necessary any more), we apply classical EM algorithm [27] to fit $P_{\mathcal{F}}$ and $P_{\mathcal{B}}$ independently on the current partition $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$. Although the algorithm works with some of the images, it generally performs less robust than the proposed local-global energy minimization algorithm. Some of the typical failure examples of the variational energy formulation without the global energy term are shown in Figure 3. The reason for the failure is that without seeking a global description of the image data, perform EM independently on the inaccurately partitioned $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$ may result in a biased estimation of $P_{\mathcal{F}}$ and $P_{\mathcal{B}}$, this will

not ensure good segmentation results.

As pointed out in [30], estimating the Gaussian mixture models in a purely unsupervised fashion can hardly keep the identity of the Gaussian component of the OOI (such as the hand in their case). Thus they proposed a restricted EM algorithm which fixes the mean of the OOI Gaussian component throughout the EM iterations. It assumes that the initial estimation is good enough which may be too strong in many practical situations. Given that we assume more general Gaussian mixture model for the OOI, the initially estimated model is usually not accurate. Thus we have to refine it during the fixed-point iterations. We do not have the identity problem because we combine the global energy with the local region energy.

2) Shape rectification: The physical shape of a business card is usually rectangle. However, in the image formation process, the rectangle shape will usually be projected as a quadrangle shape in the image. Thus the texts on the business card in the image will be skewed. It would be nice to re-transform the quadrangle shape back to a rectangle with the physical aspect ratio of the business card so we can also rectify the skewed text at the same time.

Since the business card is a planar object, it is well known that this can be easily achieved by a homography transform. It is also well known that only four pairs of correspondence points are needed to solve for a homography matrix. In fact, it is natural to choose the four corner points of the quadrangle since they are direct correspondence of the four corner points of the physical business card. To make the rectified text to look natural, at least we still need to estimate the physical aspect ratio of the business card since we have no way to obtain the physical size of the business card from a single view image. Fortunately, by making reasonable assumptions about the camera model which are easy to satisfy, if we have the image coordinates of the four corner points of the quadrangle, it has been shown in [28] that the physical aspect ratio of the rectangle can be robustly estimated given that the quadrangle is the projection of a physical rectangle shape.

Therefore, now the problem we need to address is to locate the four corner points of the quadrangle in the image. Since the segmentation subsystem returns to us the clock-wise chain code of the closed boundary of the business card, it greatly facilitate us to achieve this goal. Note that the corner points may not necessary to be on the boundary curve we obtained because it is common that one corner point be occluded by the fingers of the people who is holding it, e.g., see Figure 2(a), (b), (d), (g), (h), (i), (j), (k) and (l) for a few examples. Our solution is

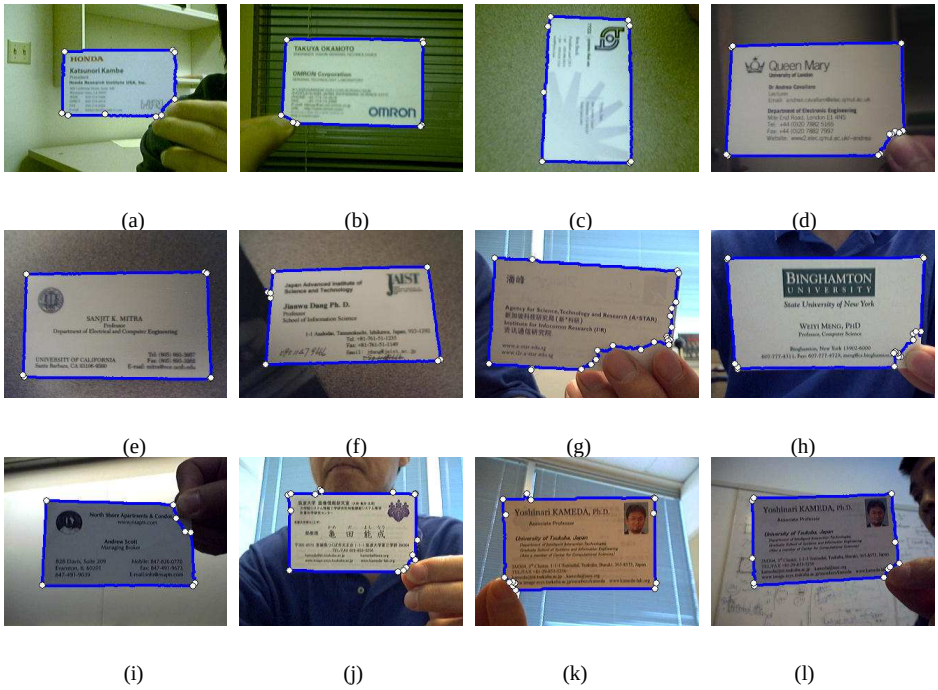


Fig. 4. Results of Curve Simplification.

to fit four lines to find a best quadrangle based on the boundary curve points and business card region. This can be achieved by the following steps.

- CURVE SIMPLIFICATION: The boundary chain code we have obtained is a dense polygon representation of the segmented area, i.e., each vertex point is in the 3×3 neighborhood of its neighboring vertex. This usually results in over 200 vertex points. As we can easily see from Figure 2, this is too redundant. Without losing much accuracy, the curve simplification procedure try to reduce the vertex to $10 \sim 20$. Denote the set of n vertex points we obtained from the segmentation subsystem as $\mathbf{V} = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\}$ with $\mathbf{v}_0$ also be the neighbor of $\mathbf{v}_{n-1}$, we perform the following two steps for curve simplification, i.e.,
  - `Multi-scale corner point detection`: Denote $(i)_m = i \ mod \ m$, for $i = 0, \ldots, n-1$, check to see if

$$\left| \frac{(\mathbf{v}_{(i-j)_n} - \mathbf{v}_i) \cdot (\mathbf{v}_{(i+j)_n} - \mathbf{v}_i)}{\|\mathbf{v}_{(i-j)_n} - \mathbf{v}_i\| \|\mathbf{v}_{(i+j)_n} - \mathbf{v}_i\|} \right| < 0.98 = \cos(10^o) \tag{37}$$

    are satisfied for all $j = 1, \ldots, m$ (we usually choose $m = 20$). If yes, we keep $\mathbf{v}_i$ in our vertex set, otherwise we remove it from the vertex set. This step in principle removes

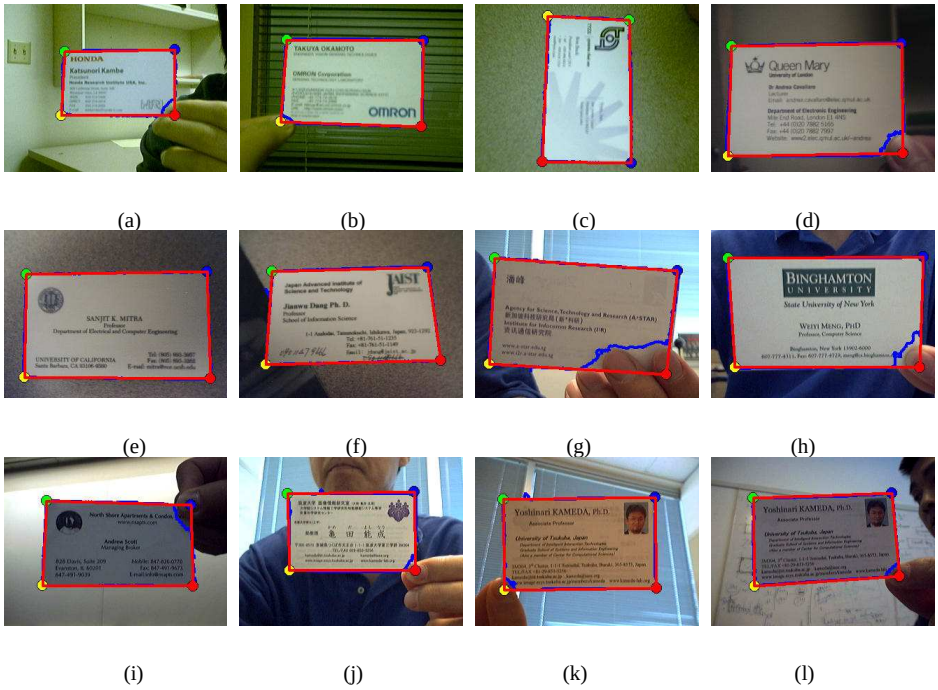


Fig. 5. Quadrangle fitting of business card.

vertex points with too small transition over multiple scales. We denote the reduced m vertex set as $\tilde{\mathbf{V}} = \{\tilde{\mathbf{v}}_0, \tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_{m-1}\}$ where $\tilde{\mathbf{v}}_0$ and $\tilde{\mathbf{v}}_{m-1}$ are again neighboring vertex.

- `Iterative minimum error vertex pruning`: For $i = 0, \ldots, m-1$, evaluate $d_i = d(\tilde{\mathbf{v}}_i, \overline{\tilde{\mathbf{v}}_{(i-1)_m}\tilde{\mathbf{v}}_{(i+1)_m}})$ the Euclidean distance from $\tilde{\mathbf{v}}_i$ to the straight line formed by its backward and forward neighbor vertices $\tilde{\mathbf{v}}_{(i-1)_m}$ and $\tilde{\mathbf{v}}_{(i+1)_m}$. Suppose $\tilde{\mathbf{v}}_k$ is such that $d_k = \min_i \{d_i\}$, if $d_k < \epsilon_d$, where ϵ_d is a pre-specified error tolerance (we usually set it to be 1), we remove $\tilde{\mathbf{v}}_k$ from $\tilde{\mathbf{V}}$. Repeat the same operations until no more vertices could be removed from the set. This returns the final reduced vertices set $\hat{V} = \{\hat{\mathbf{v}}_0, \ldots, \hat{\mathbf{v}}_{l-1}\}$.

- QUADRANGLE FITTING: We formulate the quadrangle fitting as an optimization problem. We first construct the set of all straight line candidates for the quadrangle boundary based on the pruned vertices set $\hat{V}$. We then seek the best combinations of four lines which returns the highest score according to the criterions we introduced below to obtain the best quadrangle.

(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 6. Results of rectified business card.

– `Boundary line candidate set`: For each $i = 0, \ldots, l$, construct the ordered candidate boundary line set $L_i = \{\overline{\hat{\mathbf{v}}_i\hat{\mathbf{v}}_{(i+1)_l}}, \ldots, \overline{\hat{\mathbf{v}}_i\hat{\mathbf{v}}_{(i+n_d)_l}}\} = \{l_{i1}, \ldots, l_{in_d}\}$ where n_d is an integer value which specifies how far we should look forward to form the line candidates from one specific vertex. We generally set it to be 4. We finally obtain the ordered set of all the boundary line candidates $L = \{L_1, L_2, \ldots, L_l\} = \{l_0, l_2, \ldots, l_p - 1\}$. Note that the order of the lines are also ordered according to the ordering of the vertices.
– `Quadrangle evaluation`: Denote Q_{ijkl} be the quadrangle spanned by $\{l_i, l_j, l_k, l_l\}$ where $i < j < k < l$, $\{\theta^0_{ijkl}, \theta^1_{ijkl}, \theta^2_{ijkl}, \theta^3_{ijkl}\}$ be the four corner angles spanned by the four lines. Let N_F, N_Q and $N_{F\cap Q}$ be the number of pixels identified as foreground business card pixel, the number of pixels inside the quadrangle Q_{ijkl} and the number of pixels in the intersection of the former two sets, respectively. Also, let N_c be the number of vertices point in $\mathbf{V}$ which are in the d_c neighborhood of the four line segments formed by the four lines $\{l_i, l_j, l_k, l_l\}$. We usually take n_c to be 3 pixel. Also note that there are n vertices in the vertices set $\mathbf{V}$. Then the fitness of the quadrangle

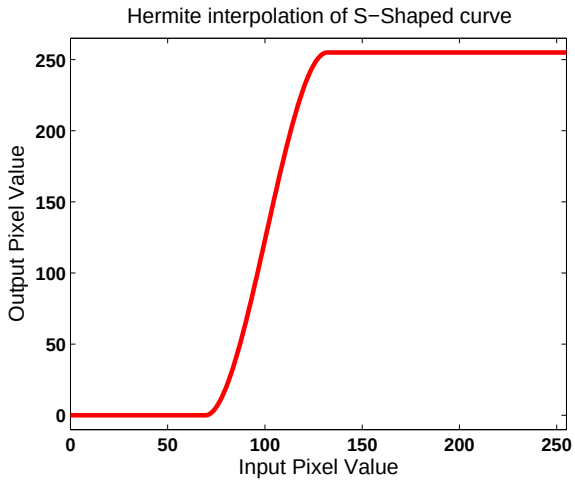


Fig. 7. An example of S-Shaped curve.

is evaluated by

$$\mathcal{S}_Q = \frac{n_c}{n}\sqrt[4]{\prod_{i=0}^{3}(1-|\cos\theta_{ijkl}^i|)}\sqrt{\frac{N_{F\cap Q}}{N_Q}+\frac{N_{F\cap Q}}{N_F}}. \quad (38)$$

The Q_{ijkl} with the largest $\mathcal{S}_Q$ is regarded as the best quadrangle. In principle the cost function $\mathcal{S}_Q$ favors the quadrangle whose boundary and enclosed region coincide with the boundary and enclosed region of the segmentation results the most. It also favors quadrangle whose corner angles are near $\frac{\pi}{2}$. This is based on the assumption that the users usually try to face the frontal of the business card to the camera. Moreover, a post-processing could be performed by collecting the Sobel edge points in the neighbor of each boundary lines and then performing a weighted least square fitting to further refine the position of each of the side lines of the quadrangle shape.

– `Heuristics to reduce computation`: The most computation intensive part of the optimization process is the evaluation of $\sqrt{\frac{N_{F\cap Q}}{N_Q}+\frac{N_{F\cap Q}}{N_F}}$ in $\mathcal{S}_Q$ since we must count the intersection of two region in the image. The following heuristics have been adopted to reduce the computation and they have been proven to be very effective:

  ∗ If the length of a line segment $\overline{\hat{\mathbf{v}}_i\hat{\mathbf{v}}_{(i+j)_l}}$, $1 \le j \le n_d$ is less than $\frac{1}{16}$ of the minimum of the image width and length, then we do not put it in the boundary line candidate set.

  ∗ If any of the corner points of the quadrangle Q_{ijkl} falls out of the image size, we

(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 8. Results of enhanced business card image.

simply discard it without evaluating the $\mathcal{S}_Q$ for it.

- If $\frac{n_c}{n} < 0.5$ for Q_{ijkl}, we simply discard it without evaluating the other term of $\mathcal{S}_Q$.
- If $|\cos\theta^i_{ijkl}| > 0.2$ for any $i = 0, \ldots, 3$, the quadrangle is discarded without further evaluation.

- RECTANGLE RECTIFICATION: Once we have obtained the best quadrangle shape Q^*_{ijkl} for the business card in the image, we can easily identify the four corner points. Then by utilizing the techniques in [28], we can easily estimate the physical aspect ratio $R_\alpha = \frac{W_l}{W_h}$ of the business card. To rectify the quadrangle, we need to determine the size of the rectangle after rectification. Since we do not want to lose any image information, i.e., each image pixel inside the quadrangle in the image must have a direct map in the rectified rectangle image, we must set the length L_r and width W_r of the rectified rectangle to be suitable to achieve this. We firstly identify the longest side of the quadrangle whose length is denoted as L_q, and denote the longer length of the two neighbor sides of the longest side as W_q. Then, if $R_\alpha W_q > L_q$, we set the $W_r = W_q$ and $L_r = R_\alpha W_q$, otherwise we set $L_r = L_q$ and $W_r = \frac{L_r}{R_\alpha}$. We then have the four corner points of the rectified rectangle to be $(0, 0)$, $(L_r, 0)$, (L_r, W_r) and $(0, W_r)$ and they are corresponding to the four corner points of the quadrangle. A homography could be estimated from the four pairs of corresponding points,

and we can easily re-warp the quadrangle shaped image patch to the rectified rectangle by reverse mapping with bi-linear color pixel interpolation.

We present the step by step results of each step of the shape rectification subsystem through Figure 4 to Figure 6.

Figure 4 presents the results of curve simplification based on the segmentation results of those images presented in Figure 2. The blue curve overlayed in each of the image is the boundary curve from our segmentation algorithm and the white points are the finally simplified vertices of the boundary curve. As we can easily observe, the curve simplification algorithm adopted significantly reduces the number of vertices of the curve while the simplified curve still represents the originally curve with high accuracy.

Figure 5 presents the results of quadrangle fitting from our optimization criterion. The green, blue, red and yellow corner points correspond to the $(0, 0)$, $(L_r, 0)$, (L_r, W_r) and $(0, W_r)$ coordinate of the rectified rectangle respectively. We also overlay the recovered quadrangle shape with red lines in the image, we can easily notice how close it is fitted with the boundary curve (blue lines) and region from our segmentation results. Also notice how the occluded corner points (mostly occluded by fingers) are recovered through quadrangle fitting, it is not a problem at all.

Figure 6 shows the results of the rectified business card image. It uses the estimated quadrangle shapes in Figure 5. Note how the skewed business card text characters are rectified at the same time with the shape rectification. Note the different aspect ratios of the rectified business card image represent very well the differences of the physical aspect ratio of these business card. For a quantitative study about the accuracy of the physical aspect ratio of the rectified rectangle shape, we refer the readers to [28].

3) Card Image Enhancement: To make the contrast of the text characters and the background in the rectified business card image more sharp, we simply independently transform the R, G, B pixel value of the rectified image through a "S" shape curve by Hermite polynomial interpolation on the average intensity $\bar{\mathcal{L}}_l$ of the lightest 10% pixels and the average intensity $\bar{\mathcal{L}}_d$ of the darkest 10% pixels. In principle, the curve should map the pixel value larger or equal to $\bar{\mathcal{L}}_l$ and pixel value less or equal to $\bar{\mathcal{L}}_d$ to near 255 and 0, respectively. Here we present in Figure 7 a "S" shaped curve interpolated on the rectified business card image in Figure 6(m).

We present the contrast enhanced business card image in Figure 8. Compared with the original

Fig. 9. Segmentation results for road sign images.

rectified business card image in Figure 6, the contrast of the color pixels has been effectively improved. However, it sometimes causes some negative effects especially when there are large light variation on the business card, e.g., Figure 8(d) (e) (f) are some examples. In this case, fitting a lighting plane like what has been utilized in [28] might be of great help.

C. Segmentation of road sign images

We have also collected a set of 37 road sign images from the internet, in which the road signs are at the focus of attention. This set of road sign images contains road signs of different shapes and different poses under a large variety of backgrounds. We have subjectively evaluated the quality of the extraction results of our algorithms by categorizing them into 3 different groups, namely “good”, “fair”, and “bad”. We have 7 different people to vote for the extraction results of each image, and the extraction result of one image is categorized to the group which it receives the largest number of votes. Overall there are 27 results being categorized as good, 5 being

(a). Fair (b). Fair (c). Fair (d). Bad (e). Bad (f). Bad

Fig. 10. Fair and bad results of road sign segmentation.

categorized as fair and 5 being categorized as bad. We present some of the sample successful results in Figure 9. As we can observe, the extraction results are quite accurate.

We also present some of the fair and bad results in Fig. 10. Two reasons may cause the unsatisfactory results: (1). The OOIs are too small or too thin in the image such as Fig. 10(e) where the tree behind the road sign is classified as the foreground object. This is because the initial OOI region contains large number of pixels of the tree. (2). There are very strong spurious edges surrounding the OOI while there is not enough differentiation between the foreground and the background colors. Fig. 10(f) is such an example where the color difference between the tree and the characters on the road sign is not strong enough 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 unsatisfactory results for extracting general OOIs in Section V-D.

D. Segmentation of other objects

To test the ability and robustness of the proposed algorithm to extract general object of interest from static images, we have tested the proposed algorithm on a set of 63 images, in which the OOI is at the focus of attention, from the Berkley image database [31]. This set of images are more challenging because the appearances of the OOIs and the background are more complex. With the same subjective evaluate method as that for the road sign image set, the extraction results on 31 images are categorized as good, 15 are categorized as fair and 17 are categorized as bad. We present some typical successful extraction results on this image database in Figure 11.

With the same setting as the segmentation algorithm in Section V-B, we have also obtained successful segmentation results in a variety animal images such as dog, wolf, rabbit, squirrel, zebra, raccoon and hawk downloaded from internet. We also obtained successful results on

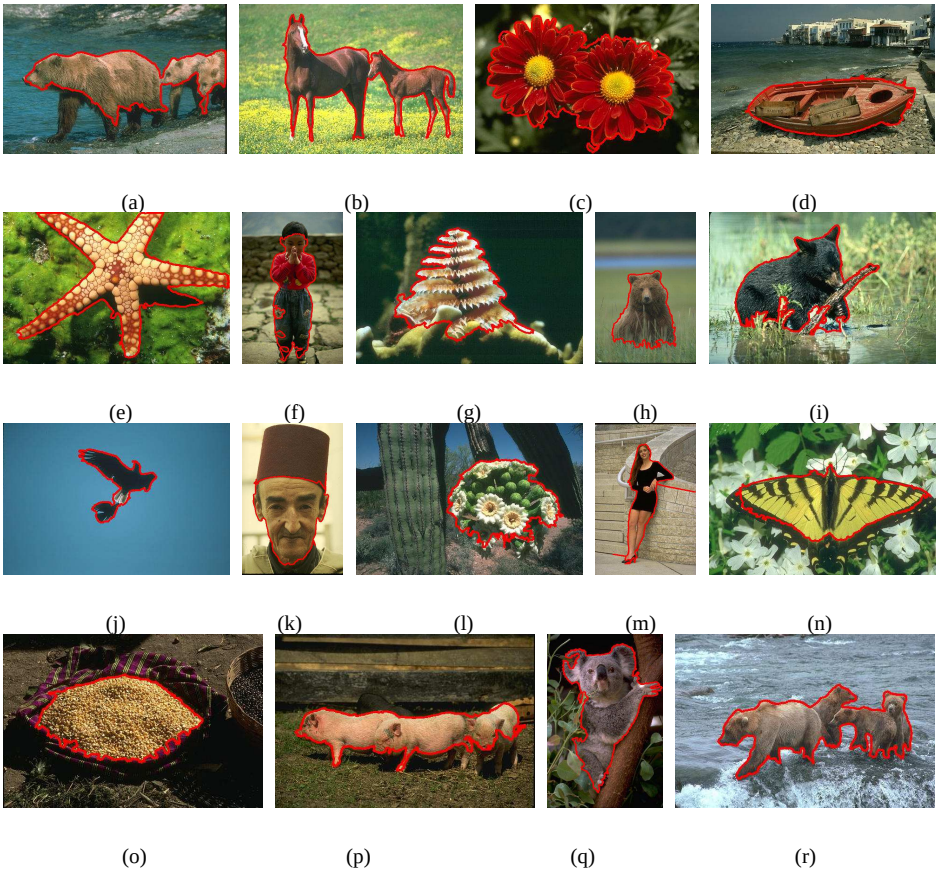


Fig. 11. Segmentation results of general objects on the Berkeley image data-base.

segmenting human hand and head, cell phone and telephone, cups and even receipts. We present some of the results in Figure 12. The results are quite accurate.

E. Validation of the quasi-semi-supervised EM on real experiments

One may have the concern that why the local-global variational energy formulation can achieve better results. The reason is that the quasi-semi-supervised EM algorithm generally will achieve more accurate estimation of the joint data likelihood model and thus more accurate estimation of the foreground and background model, since the objective function incorporated a term to maximize the joint data likelihood. In fact, this can be demonstrated by analyzing the failure case of the variational energy formulation without the data likelihood potential in Figure 3(c). We plot the marginal foreground background distribution obtained from both formulations in Figure 13. For convenience, we call the variational formulation without the global data likelihood potential

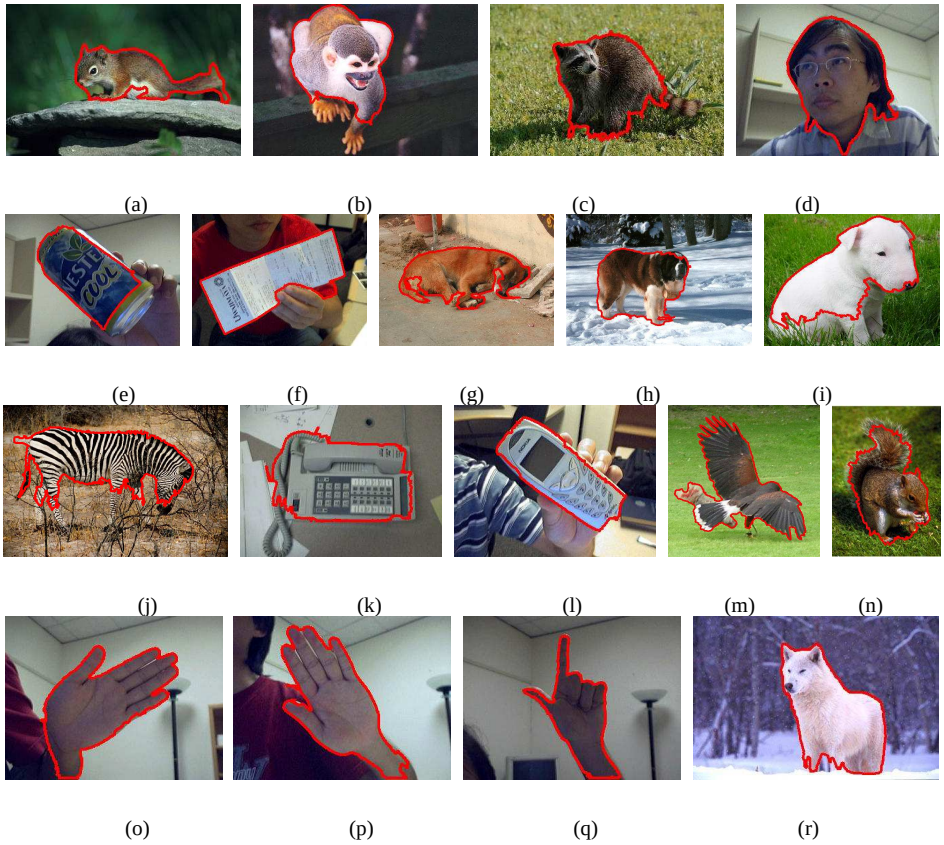


Fig. 12. Segmentation results of other general objects.

as local variational formulation.

From left to right, the first row in Figure 13 presents the estimated marginal foreground/background distributions by our algorithm upon convergence on the L, U and V dimension, respectively. The second row of Figure 13 shows the ground-truth marginal distribution on L, U and V, which are estimated on the manually annotated foreground/background on the image shown in Figure 3(c) (actually, different intermediate results of the same image are shown in (h) of Figure 2, 4 and 5, respectively.). The third row presents the three marginal distributions estimated by the algorithm deduced on the local variational formulation.

It is worthwhile to mentioning how close are the marginal distributions estimated by our algorithm to the ground-truth marginal distributions. In contrast, also note that how far away are the foreground/background distributions obtained by the local variational formulation to those of the ground-truth . It is obvious that the failure of the local variational formulation is due to

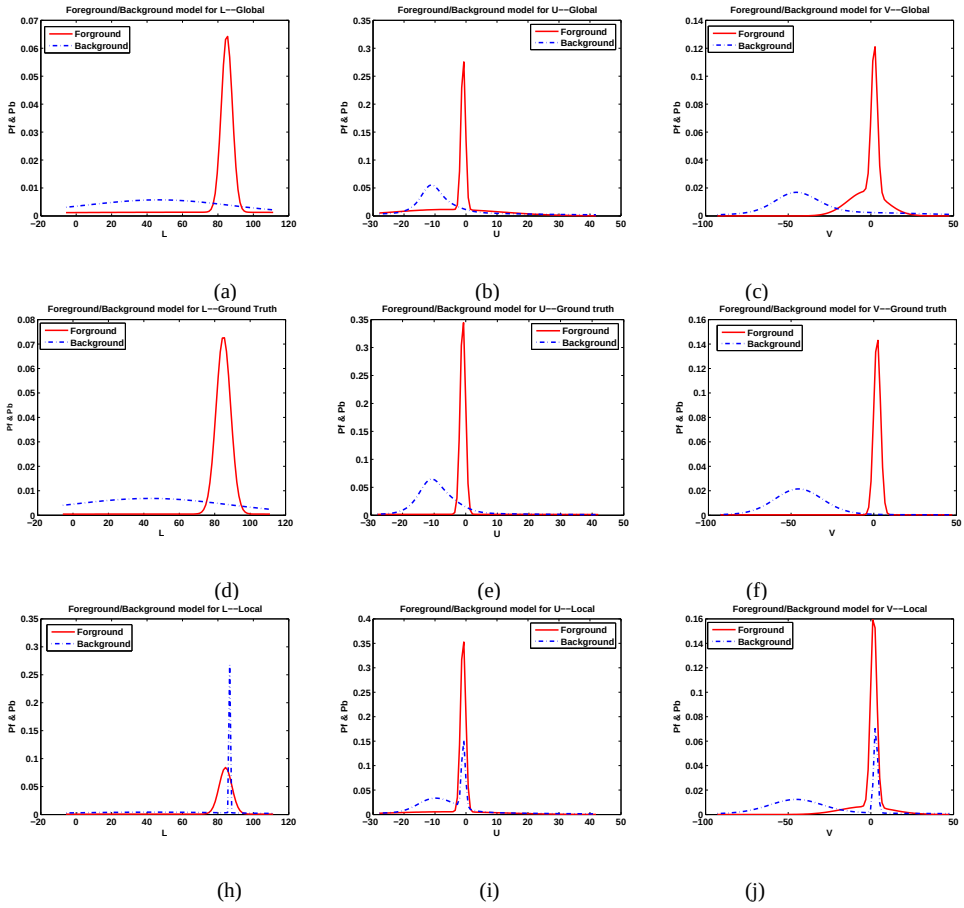


Fig. 13. The estimated marginal distribution of foreground (red line) and background (blue line) on L (first column), U (second column) and V (last column). The first row presents these distributions obtained by our algorithm upon convergence. The second row presents the ground-truth distributions estimated on manually labelled foreground/background regions. And the third row presents these distributions estimated by local variational formulation. The comparison is performed on the image shown in Figure 3 (c).

the inability to accurately estimate the foreground/background distributions. This is not strange because under the local variational formulation, the two distributions are independently estimated by traditional EM algorithm [27] over the region $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{A}_{\mathcal{B}}$, respectively, which are doomed by the local fitting problem. On the other hand, our fixed-point iterations nicely solved this problem since it also maximize the global image data likelihood at the same time.

VI. Conclusion and Future Work

This paper proposes a novel local-global variational energy formulation for image segmentation, based on which an iterative scheme is formulated to perform the minimization of the energy.

Our main contributions are (1) the incorporation of a global image data likelihood potential to better estimate the foreground/background distributions, and (2) a set of fixed-point equations which we call quasi-semi-supervised EM for Gaussian mixture models. As we have discussed, the quasi-semi-supervised EM is specially suited to deal with the learning problem with inaccurate labeled data where an unknown portion of the labels of the data are erroneous.

Based on the proposed approach, we have built a real time system to segment, rectify and enhance business card images. Our formulation and algorithm are also general to segment other general objects. Extensive experiments have demonstrated the effectiveness and efficiency of the proposed approach. Future work includes extending the variational energy formulation for the segmentation of multiple objects.

REFERENCES

[1] S. C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation," *IEEE Transaction on Pattern Recognition and Machine Intelligence*, vol. 18, no. 9, pp. 884–900, 9 1996.

[2] N. Paragios and R. Deriche, "Geodesic active regions and level set methods for supervised texture segmentation," *International Journal of Computer Vision*, pp. 223–247, 2002.

[3] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut"– interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (SIGGRAPH'04)*, pp. 309–314, 2004.

[4] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321–331, 1987.

[5] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problem," *Communications on Pure and Applied Mathematics*, vol. 42, pp. 577–684, 1989.

[6] L. D. Cohen and I. Cohen, "Finite-element methods for active contour models and balloons for 2-d and 3-d images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1131–1147, 11 1993.

[7] V. Casselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[8] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transaction on Image Processing*, vol. 10, no. 2, pp. 266–277, Feburary 2001.

[9] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulation," *Journal of Computational Physics*, vol. 79, pp. 12–49, 1988.

[10] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: a level set approach," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 158–175, Feburary 1995.

[11] D. Peng, B. Merriman, S. Osher, H. Zhao, and M. Kang, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulation," *Journal of Computational Physics*, vol. 155, pp. 410–438, 1999.

[12] C. Li, C. Xu, C. Gui, and M. D. Fox, "Level set evolution without re-initialization: A new variational formulation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, San Diego, June 2005, pp. 430–436.

[13] Y. Shi and W. C. Karl, "Real-time tracking using level sets," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, San Diego, June 2005, pp. 34–41.

[14] J.-F. Aujol and G. Aubert, "Signed distance functions and viscosity solutions of discontinuous hamilton-jacobi equations," INRIA, Tech. Rep. RR-4507, 2002.

[15] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 721–741, June 1984.

[16] O. Veksler, "Efficient graph-based energy minimization methods in computer vision," Ph.D. dissertation, Cornell University, 1999.

[17] Y. Boykov and M.-P. Jolly, "Interactive organ segmentation using graph cuts," in *Proc. of International Society and Conference Series on Medical Image Computing and Computer-Assisted Intervention*, vol. LNCS 1935, 2000, pp. 276–286.

[18] Y. Boykov, V. S. Lee, H. Rusinek, and R. Bansal, "Segmentation of dynamic n-d data sets via graph cuts using markov models," in *Proc. of International Society and Conference Series on Medical Image Computing and Computer-Assisted Intervention*, vol. LNCS 2208, 2001, pp. 1058–1066.

[19] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *8th IEEE International Conference on Computer Vision*, vol. 1, July 2001, pp. 105–112.

[20] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, September 2004.

[21] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Feburary 2004.

[22] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1222–1239, November 2001.

[23] A. Barbu and S.-C. Zhu, "Graph partition by swendsen-wang cut," in *Proc. IEEE International Conference on Computer Vision*, 2003.

[24] ——, "Generalizing swendsen-wang to sampling arbitrary posterior probabilities," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, 2005.

[25] Z. Tu, "An integrated framework for image segmentation and perceptual organization," in *Proc. of IEEE International Conference on Computer Vision*, Beijing, China, Octobor 2005.

[26] N. Paragios and R. Deriche, "Geodesic active contours for supervised texture segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 1999, pp. 1034–1040.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[28] Z. Zhang and L. He, "Notetaking with a camera: Whiteboard scanning and image enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Montreal, Quebec, Canada, May 2004, pp. 533–536.

[29] D. Comaniciu and P. Meer, "Mean-shift: A robust approach toward feature space analysis," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 1–18, May 2002.

[30] X. Zhu, J. Yang, and A. Waibel, "Segmenting hands of arbitrary color," in *Proc. IEEE International Conference on Automatic Face Recognition*. Grenoble, France: IEEE Computer Society, March 2000, pp. 446–453.

[31] 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," in *8th IEEE International Conference on Computer Vision*, vol. 2, July 2001, pp. 416–423.

Gang Hua is a Ph.D. candidate in the Department of Electrical Engineering and Computer Science, Northwestern university. His main research interests include Computer Vision, Computer Graphics, Visual Tracking and Machine Learning. During the summer 2005 and 2004, he was a research intern with the Speech Technology Group, Microsoft Research, Redmond, Washington, and a research intern with the Honda Research Institute, Mountain View, California, respectively. Before attending in Northwestern in 2002, He was a master student in the AI&R Institute in Xi'an Jiaotong University(XJTU), Xi'an, P.R.China. He received his M.S. in Control Science and Engineering at XJTU in 2002. He was enrolled in the Special Class for the Gifted Young of XJTU in 1994 and received his B.S. in Automatic Control Engineering in 1999.

He received the Richter Fellowship and the Walter P. Murphy Fellowship at Northwestern University in 2005 and 2002, respectively. When he was in XJTU, he was awarded the Guanghua Fellowship, the EastCom Research Scholoarship, the Most Outstanding Student Exemplar Fellowship, the Sea-star Fellowship and the Jiangyue Fellowship in 2001, 2000, 1997, 1997 and 1995 respectively. He was also a recipient of the University Fellowhip for Outstanding Student of XJTU from 1994 to 2002.

PLACE
PHOTO
HERE

Zicheng Liu Biography text here.

PLACE
PHOTO
HERE

Zhengyou Zhang Biography text here.

Ying Wu received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 1994, the M.S. degree from Tsinghua University, Beijing, China, in 1997, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001.

From 1997 to 2001, he was a Graduate Research Assistant at the Image Formation and Processing Group of the Beckman Institute for Advanced Science and Technology at UIUC. During summer 1999 and 2000, he was a Research Intern with the Vision Technology Group, Microsoft Research, Redmond, Washington. Since 2001, he had been on the faculty of the Department of Electrical and Computer Engineering at the Northwestern University, Evanston, Illinois. His current research interests include computer vision, computer graphics, machine learning, human-computer intelligent interaction, image/video processing, multimedia, and virtual environments. He received the Robert T. Chien Award at the University of Illinois at Urbana-Champaign in 2001, and is a recipient of the NSF CAREER award.