Technical Report MSR-TR-2006-10

Automatic Segmentation of Objects of Interest from an Image

Gang Hua[†], Zicheng Liu[‡], Zhengyou Zhang[‡], Ying Wu[†]
[†]2145 Sheridan Road, EECS Dept.
Northwestern University
Evanston, IL 60208, U.S.A.
{ganghua, yingwu}@ece.northwestern.edu
[‡]One Microsoft Way
Microsoft Research
Redmond, WA 98052, U.S.A.
{zliu, zhang}@microsoft.com

January 27, 2006

Abstract

In this paper, we propose a novel variational energy formulation for image segmentation. Traditional variational energy formulation for image segmentation like that in [41, 30] only incorporates local region potentials. 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 iterative energy minimization process. 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 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 business card extraction, road sign extraction and general object-of-interest segmentation, demonstrate the robustness, effectiveness, and efficiency of the proposed approach.

Contents

1 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 and background models, we can directly separate the OOI from the background like what is done in the geodesic active region approach [30]. 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 iterative energy minimization problem, i.e., variational energy minimization [41, 34] or graph energy minimization [33]. The coherent image regions are usually modeled in a probabilistic way as Gaussian distributions [41], Gaussian mixture models (GMM) [30, 5, 33], or even kernel densities [34].

One problem with the iterative energy minimization 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 ([41, 30, 34, 33]) 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 2; the detailed discussion of our variational energy formulation with the global image likelihood potential is presented in Section 3; in Section 4, 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 5, 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 6.

2 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 [20]. Later work include the Mumford-Shah model [27, 37], the active contour with balloon forces [13], the region competition algorithm [41], the geodesic active contours [11, 31], region based active contour [1, 12, 18, 21], the geodesic active region [30],and active contour with shape derivatives [19], etc. The energy functionals constructed in this track are usually formulated on the region boundary curves [20, 13, 11] and/or over the regions partitioned by the boundary curves [41, 12, 30].

In practice, energy functionals based purely on image gradient information like what was proposed in [20, 13, 11, 37, 31] are easy to get stuck in a local optima especially when there are many spurious edges in the image. On the other hand, using or combining the intensity, color and texture distributions [41, 12, 30, 35, 18, 19, 21] of the image pixels over the regions to formulate the energy functional can largely overcome this problem. Region based energy formulation can be categorized into two: *supervised* method [29, 30, 1, 35, 19] and *unsupervised* method [41, 34, 12, 18, 21]. 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 the energy with respect to the region boundary and region models alternatively. The methods of minimizing the energy with respect to the region boundary has evolved from the traditional finite difference method (FDM) [20, 41] and the finite element method (FEM) [13] to the more advanced level set method [28, 25, 11, 12, 30, 18, 19, 21].

There are a lot of works 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 [32], the level set without re-initialization [24] and the fast level set implementation without solving PDEs [36]. In fact, all these efficient level set algorithms take advantage of the property of the signed distance function [2], which is usually adopted as the implicit level set surface [30, 24, 36].

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) [17, 39]. A lot of successful results have been proposed in recent years such as the interactive object extraction [6, 9, 7] and the iterative Grab-cut system [33], where an efficient min-cut/max-flow algorithm proposed in [8] 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 [22]. For energy functions with multi-label variables, approximate solution can be obtained by using the algorithm proposed in [10] 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 [8], the types of energy functions it can minimize are still limited [22]. 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 [3, 4] and the generalized m-way Swendsen-Wang cut [38].

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 [20, 41], FEM [13] and level set [30], while the graph energy minimization could be solved by min-cut/max-flow algorithms such as the one in [8] and the Swendsen-Wang cut [3, 4, 38]. 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.

3 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 such as Gaussian distributions [41], Gaussian mixture models (GMM) [30, 5, 33], or even kernel densities [34].

Our goal is to extract a foreground object from the background. It is not realistic to assume the foreground object or the background region be 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

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

where π_i, μ_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.

3.1 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 [41, 30] 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 [41, 30].

3.2 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. This problem arises because we ignore the optimality of the global description of the whole images if we only maximize Equation

4. Therefore, we propose to add a global image likelihood potential that seeks for a global description of the entire image data. The intuition is that seeking for the global description at the same time may largely reduce the negative effects of those erroneously labeled pixels.

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

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

Later in the experiments we will further 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 strong cues for segmentation. There has been a significant literature which incorporate edge information into a variational energy function such as the SNAKES [20], the active contour model with balloons [13] and the geodesic active contour [29], 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.

3.4 Boundary, region and data likelihood synergism

With all the above energy terms, our energy formulation is the synergism of the three, i.e.,

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

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 [41] and the geodesic active region [30], the uniqueness of our formulation is that none of their formulation incorporated the *global image data likelihood potential*. Moreover, the region competition approach [41] 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 [30], the foreground and background distributions are pre-trained which renders the potential energy to be only dependent on the boundary curve $\Gamma(c)$.

3.5 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 problem statement.

Problem 1 *Let $\mathcal{D} = \{d_i | 1 = 1 \dots n\}$ 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 $\mathcal{L} = \{l_i | l_i \in \{1, 2\}, i = 1 \dots n\}$ 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 $\mathcal{Z} = \{z_i | z_i \in \{1, 2\}, i = 1 \dots n\}$ 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 [16].

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 above in Problem 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.,

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

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.

4 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.

4.1 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

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

where $\mathbf{n}(\cdot)$ 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 [30] 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 [30, 29], 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 $\Gamma(c,t+\tau) = \{(x,y)|\varphi(x,y,t+\tau)=0\}$.

4.2 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.,

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

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

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

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

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

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

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

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

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

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

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.,

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

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

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

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

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

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

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

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

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 3.5. In [23], 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.

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 5.

5 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

	$\beta = 0.01$	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.15$	$\beta = 0.2$	$\beta = 0.25$	$\beta = 0.5$
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	5.4 ± 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	5.6 ± 3.3
Better LGEM (%)	100%	99.9%	99.3%	87.6%	72.1%	57.2%	34.2%

Table 1: Comparison of local-global energy minimization (LGEM) and local energy minimization (LEM). The table shows the average joint KLs distance and its standard deviation between the estimated model and the ground truth from 1000 simulations for each β.

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.

5.1 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

$$\begin{aligned} P(\mathbf{d}|\mathbf{\Theta}) &= \omega_1 P_1(\mathbf{d}|\mathbf{\Theta}_1) + \omega_2 P_2(\mathbf{d}|\mathbf{\Theta}_2) \\ &= \omega_1 \left(\pi_{11}\mathcal{N}(\mathbf{d}|\mu_{11}, \sigma_{11}^2) + \pi_{12}\mathcal{N}(\mathbf{d}|\mu_{12}, \sigma_{12}^2)\right) \\ &+ \omega_2 \left(\pi_{21}\mathcal{N}(\mathbf{d}|\mu_{21}, \sigma_{21}^2) + \pi_{22}\mathcal{N}(\mathbf{d}|\mu_{22}, \sigma_{22}^2)\right), \end{aligned} \tag{36}$$

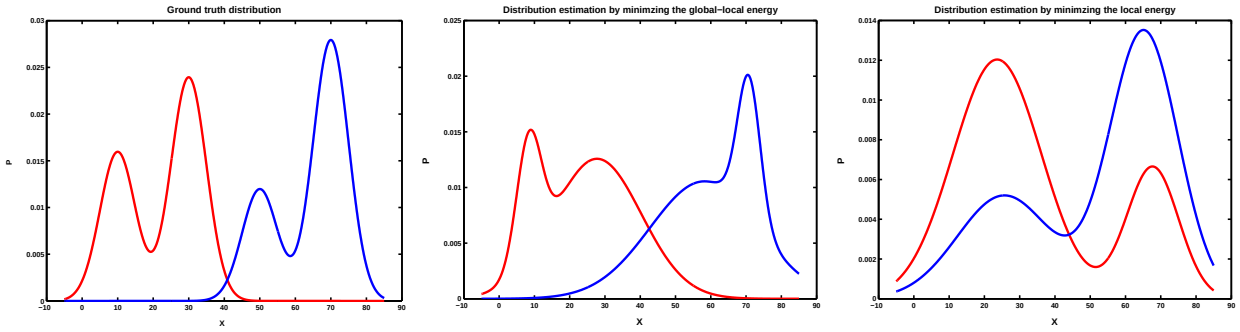
where $\mathcal{N}(\mathbf{d}|\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2. Then $\mathbf{\Theta}$ represents the set of parameters for all the Gaussian mixture components. With a specific $\mathbf{\Theta}$, we randomly draw a set $\mathcal{D}$ of 20000 data samples 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% 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-semi-supervised EM algorithm similar to that in Section 4.2. The local energy minimization (LEM) is performed by applying the classical EM algorithm [16] 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 respect to the ground truth by the following *joint KLs distance*, i.e.,

$$\mathcal{D}(P^*, P) = KL_s(\omega^*, \omega) + KL_s(P_1^*, P_1) + KL_s(P_2^*, P_2) \tag{37}$$

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 Equation 37 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 1000 data models and thus run 1000 simulations of the experiments described above. The experimental results are listed in Table 1. For both algorithms, in each simulation we randomly choose 10 different initializations and the best results are adopted.



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

Figure 1: Visual comparison of model estimation by local and local-global energy minimization in one simulation ($\beta = 0.05$) .

Table 1 clearly shows that with 30% 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 1000 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, i.e., when $\beta > 0.2$, the performance of LGEM is almost the same or even worse than LEM. The third row of Table 1 presents the percentage of the 1000 simulations for a fixed β where the LGEM estimated better model than the LEM. It also clearly shows that LGEM will degrade with the increase of β.

We observed similar results for 20% and 10% erroneous labels, although β needs to be larger to degrade the 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-semi-supervised EM may take much longer to converge. Following the idea of [15], theoretic analysis of choosing β may be possible, we defer it to our future work. As pointed out in [42], 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 Figure 1 the models estimated from LGEM, from LEM, and the ground truth in one simulation when $\beta = 0.05$. As we can observe, the model from LGEM is far more accurate than that from LEM.

5.2 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 [40]. 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.

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

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

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

Figure 2: Card segmentation results.

5.2.1 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 [14]. 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.

(a) (b) (c)

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

- 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 $\mu_1^{\mathcal{F}}$ and $\mu_2^{\mathcal{F}}$. 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 $\mu_i^{\mathcal{B}}$ 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 $\frac{1}{8}$.
- 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 [16] 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 [42], 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.

5.2.2 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 homograph transform. It is also well known that only four pairs of correspondence points are needed to solve for a homograph 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 [40] 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 to fit four lines to find a best quadrangle based on the boundary curve points and business card region. The output from the segmentation algorithm is the business card boundary chain code, which is a dense polygon representation. We denote $\mathbf{V} = \{\mathbf{v}_0, \ldots, \mathbf{v}_{n-1}\}$ as the set of n dense vertices, and denote $\mathcal{L} = \{l_0, \ldots, l_{r-1}\}$ be the $r = C_n^2$ line sets formed by all lines $\overline{\mathbf{v}_i\mathbf{v}_j}$, where $i \neq j$ and $0 \leq i, j \leq n-1$. The set of all s quadrangles formed by four different lines in $\mathcal{L}$ is denoted by $\mathcal{Q} = \{Q_0, \ldots, Q_{s-1}\}$ where $s = C_r^4$. We also denote $\{\theta^0, \theta^1, \theta^2, \theta^3\}$ as the four corner angles of quadrangle $Q \in \mathcal{Q}$. Let N_F, N_Q and $N_{F \cap Q}$ be the three numbers of pixels in the segmented card region, inside Q and in the intersection of the former two, respectively. Also let n_c be the number of vertices in $\mathbf{V}$ which are in the 3×3 neighborhood of the boundary line

pixels of Q. The fitness of the Q to the segmented region is defined as

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

Therefore, the quadrangle fitting is just to solve for an optimization problem which seeks for the best Q^* such that

$$Q^* = \arg\max_{Q\in\mathcal{Q}} \mathcal{S}(Q). \tag{39}$$

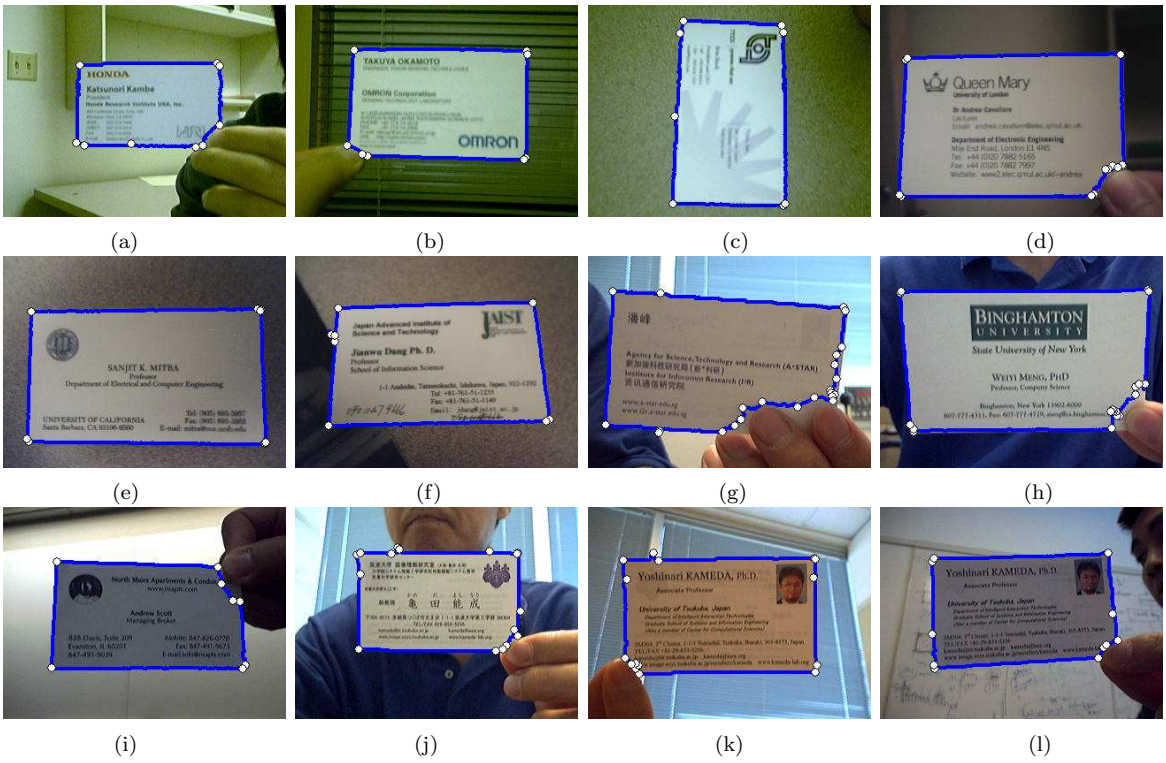
In principle, it favors quadrangles whose boundary and enclosed region coincide the most with the segmented business card image region. It also favors quadrangles whose corner angles are near $\frac{\pi}{2}$. This is based on the assumption that the users try to face the front of the business card to the camera.

However, when $n = 300$ (typical for $\mathbf{V}$), the cardinality of $\mathcal{Q}$ is $s = C_r^4 = 4.024e + 016$. It is infeasible to exhaustively search in $\mathcal{Q}$ for Q^*. We develop the following approach to pruning the solution space, and thus obtain the Q^* more efficiently.

- Curve Simplification: Indeed $\mathbf{V}$ is a highly redundant shape representation and can be simplified with high accuracy by:
  - `Multiscale corner point detection`: Denote $(i)_m = i \ mod \ m$, then for $i = 0,\ldots,n-1$, check whether

$$\left|\frac{(\mathbf{v}_{(i-j)_n}-\mathbf{v}_i)\cdot(\mathbf{v}_{(i+j)_n}-\mathbf{v}_i)}{\|\mathbf{v}_{(i-j)_n}-\mathbf{v}_i\|\|\mathbf{v}_{(i+j)_n}-\mathbf{v}_i\|}\right| < 0.98 = \cos(10^o) \tag{40}$$

    are satisfied for all $j = 1,\ldots,q$ (e.g., $q = 20$). If yes, we keep $\mathbf{v}_i$, otherwise we remove it. This step in principle removes vertices with too small transitions over multiple scales. We denote the reduced m vertices set as $\tilde{\mathbf{V}} = \{\tilde{\mathbf{v}}_0, \tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_{m-1}\}$.
  - `Iterative minimum distance pruning`: For each $i = 0,\ldots,m-1$, 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 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 tolerance (e.g., $\epsilon_d = 1$), we remove $\tilde{\mathbf{v}}_k$ from $\tilde{\mathbf{V}}$. Repeat the same operations until no more vertices could be removed. This returns the final reduced l vertices set $\hat{V} = \{\hat{\mathbf{v}}_0,\ldots,\hat{\mathbf{v}}_{l-1}\}$.
- Quadrangle pruning: With the pruned vertices set $\hat{V}$, we further prune the number of lines and quadrangles:
  - `Ordered Lines`: For each vertex $\tilde{\mathbf{v}}_i$ where $0 \le i \le l-1$, build the ordered line set $\mathbf{L}_i = \{\overline{\hat{\mathbf{v}}_i\hat{\mathbf{v}}_{(i+1)_l}},\ldots,\overline{\hat{\mathbf{v}}_i\hat{\mathbf{v}}_{(i+n_d)_l}}\} = \{l_{i1},\ldots,l_{in_d}\}$ where n_d specifies how far (e.g., $n_d = 4$) to look forward to form the lines. We obtain the ordered set $\mathbf{L} = \{\mathbf{L}_1,\ldots,\mathbf{L}_l\} = \{l_0,\ldots,l_{p-1}\}$. Notice that placing the order constraint reduces the number of lines from C_l^2 to $l \cdot n_d$.
  - `Quadrangle searching`: Let $\mathbf{Q} = \{Q_{ijkf} : 0 \le i < j < k < f \le p-1\}$ be all the possible quadrangles Q_{ijkf} spanned by $\{l_i, l_j, l_k, l_f\}$. Notice that not all quaternion $\{l_i, l_j, l_k, l_f\}$ can span a quadrangle. We then exhaustively search for the best quadrangle Q^* on $\mathbf{Q}$, i.e., $Q^* = \arg\max_{Q\in\mathbf{Q}}\mathcal{S}(Q)$. Moreover, a post-processing is performed, by first collecting the Sobel edge points in the neighborhood of the boundary line pixel of Q^* and then performing a weighted least square fitting to further refine the position of each side line of Q^*.



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

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

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

Figure 4: Results of Curve Simplification.

– `More heuristics`: The most intensive computation is to evaluate $\sqrt{\frac{N_{F\cap Q}}{N_Q}+\frac{N_{F\cap Q}}{N_F}}$ in $\mathcal{S}(Q)$, since we must count the number of pixels in the intersection of two image regions. The following criterions have been proven to be very effective to reduce the computation without losing much accuracy:
  * If the segment length of $\overline{\hat{\mathbf{v}}_i\hat{\mathbf{v}}_{(i+j)_l}}$, $1 \le j \le n_d$ is less than $\frac{1}{16}$ of the minimum of the image width and length, then we do not put it in $\mathbf{L}_i$.
  * If any corner point of Q_{ijkf} falls out of the size of the image, we simply discard it.
  * If $\frac{n_c}{n} < 0.5$ for Q_{ijkf}, we simply discard it.
  * If $|\cos\theta^w| > 0.2$ for any $w = 0, \ldots, 3$, the quadrangle Q is discarded without further evaluation.

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

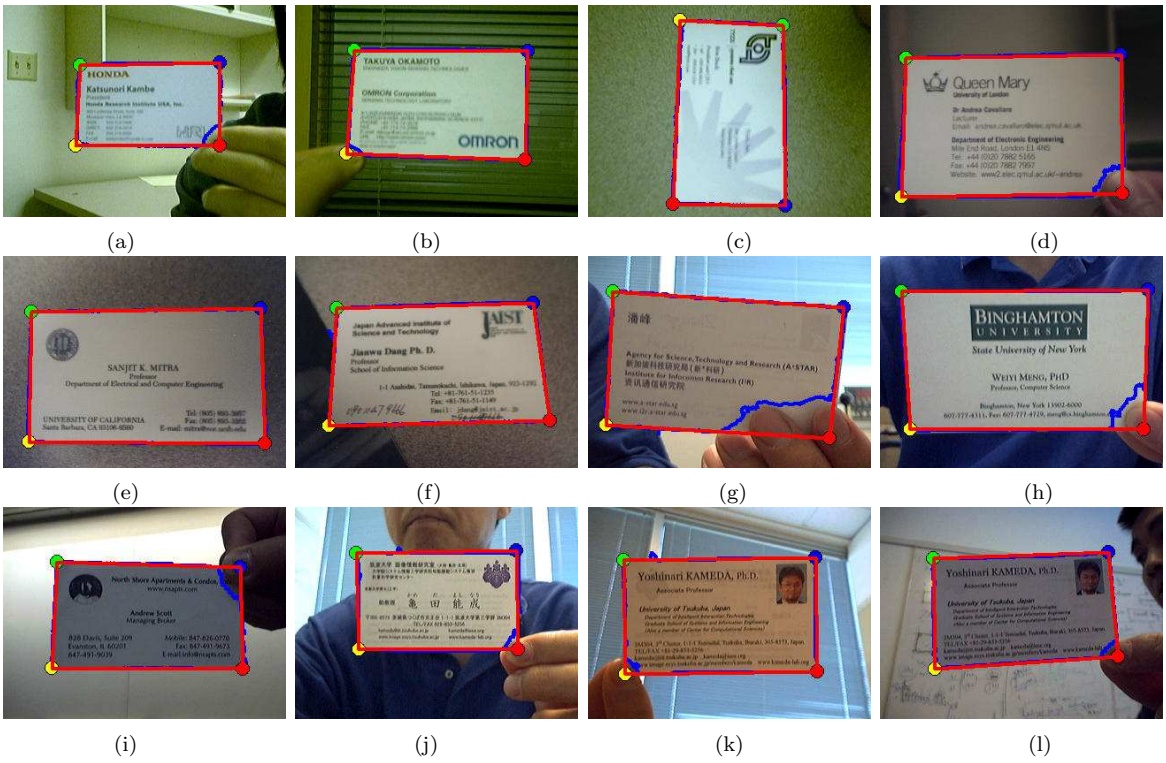


Figure 5: Quadrangle fitting of business card.

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 [40].

5.2.3 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

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

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

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

Figure 6: Results of rectified business card.

the rectified business card image in Figure 6(m).

We present the contrast enhanced business card image in Figure 8. Compared with the original 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 [40] might be of great help.

5.3 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 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 Figure 10. Two reasons may cause the unsatisfactory results: (1). The OOIs are too small or too thin in the image such as Figure 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. Figure 10(f) is such an example where the color difference between the

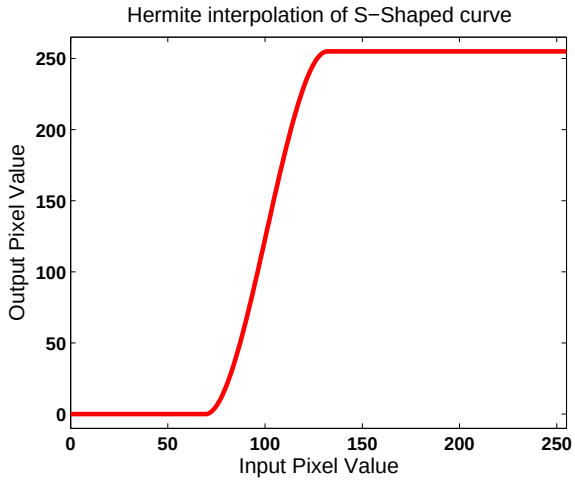


Figure 7: An example of S-Shaped curve.

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 5.4.

5.4 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 [26]. 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 5.2, 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 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.

5.5 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 [16] 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]

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

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

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

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

Figure 8: Results of enhanced business card image.

	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

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.

are summarized in Table 2. As we can notice, over the 37 road sign images, the local energy formulation obtains 19 good, 7 fair and 8 bad results, which are significantly inferior to the 27 good, 5 fair and 5 bad results from our local-global energy formulation. While over the 63 Berkeley images, the local energy formulation obtains 19 good results, 16 fair results and 28 bad results, which is again significantly inferior to the results obtained from 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 obtains a good result, our approach can also obtain a good result on the same image. On the other hand, on a significant number of test images, our approach obtains good results and the local energy formulation failed to achieve that. It is the global image likelihood potential makes the difference, because it enables the model estimation step to be more accurate.

To further demonstrated it, it is easy to analyze the failure case of the variational energy for-

Figure 9: Segmentation results for road sign images.

mulation 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 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

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

Figure 10: Fair and bad results of road sign segmentation.

those of the ground-truth . It is obvious that the failure of the local variational formulation is due to 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 [16] 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.

6 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] Jr Anthony Yezzi, Andy Tsai, and Alan S. Willsky. A statistical approach to snakes for bimodal and trimodal imagery. In *Proc. of IEEE International Conference on Computer Vision*, pages 898–903, Kerkyra, Corfu, Greece, September 1999.

[2] Jean-François Aujol and Gilles Aubert. Signed distance functions and viscosity solutions of discontinuous hamilton-jacobi equations. Technical Report RR-4507, INRIA, 2002.

[3] Adrian Barbu and Song-Chun Zhu. Graph partition by swendsen-wang cut. In *Proc. IEEE International Conference on Computer Vision*, 2003.

[4] Adrian Barbu and Song-Chun Zhu. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(8), 2005.

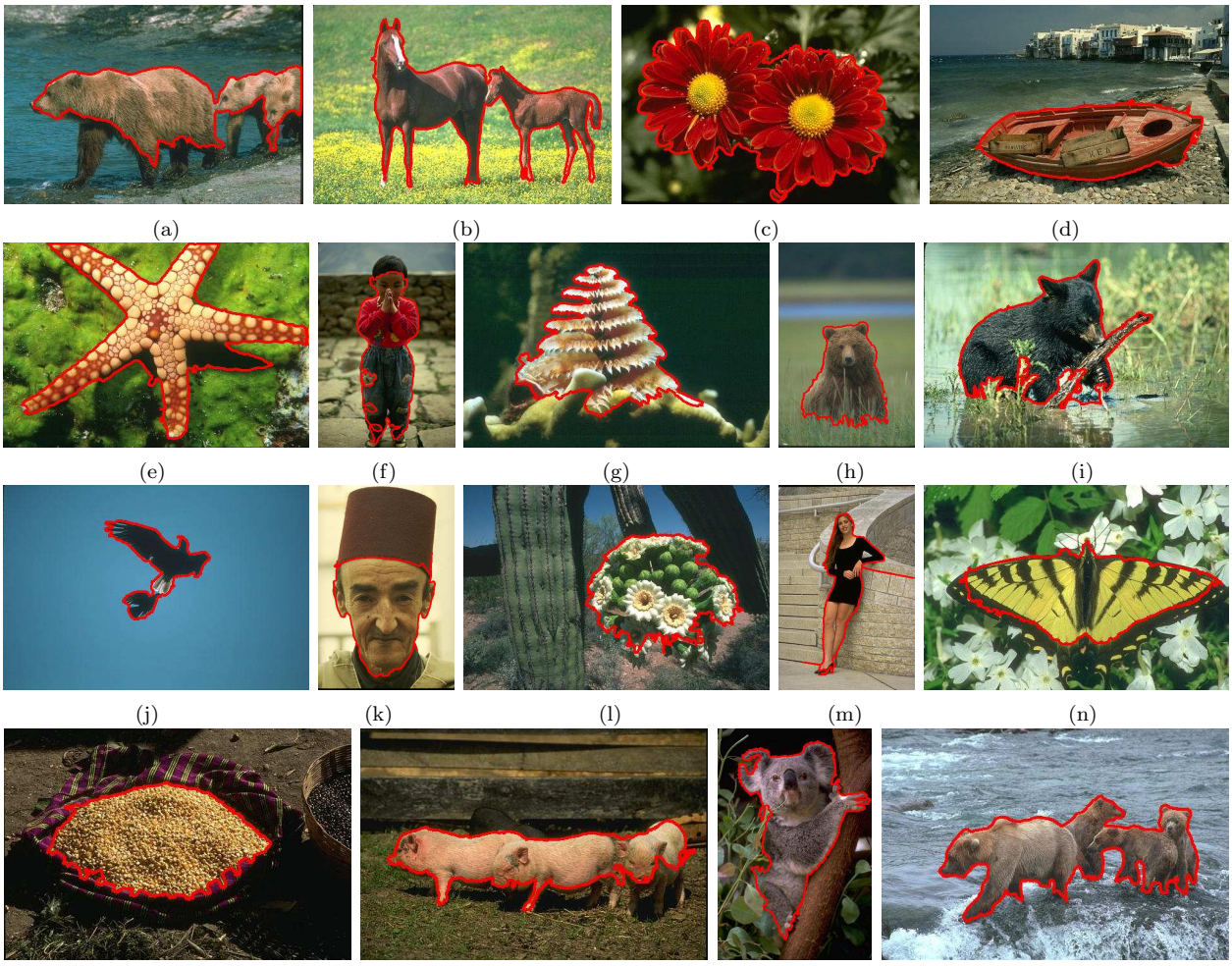


Figure 11: Segmentation results of general objects on the Berkeley image data-base.

[5] Andrew Blake, Carsten Rother, Matthew Brown, Patrick Pérez, and P. Torr. Interactive image segmentation using an adaptive gaussian mixture mrf model. In *Proc. of the 8th European Conference on Computer Vision*, pages 428–441, Prague, Czech Republic, 2004.

[6] Yuri Boykov and Marie-Pierre Jolly. Interactive organ segmentation using graph cuts. In *Proc. of International Society and Conference Series on Medical Image Computing and Computer-Assisted Intervention*, volume LNCS 1935, pages 276–286, 2000.

[7] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *8th IEEE International Conference on Computer Vision*, volume 1, pages 105–112, July 2001.

[8] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, September 2004.

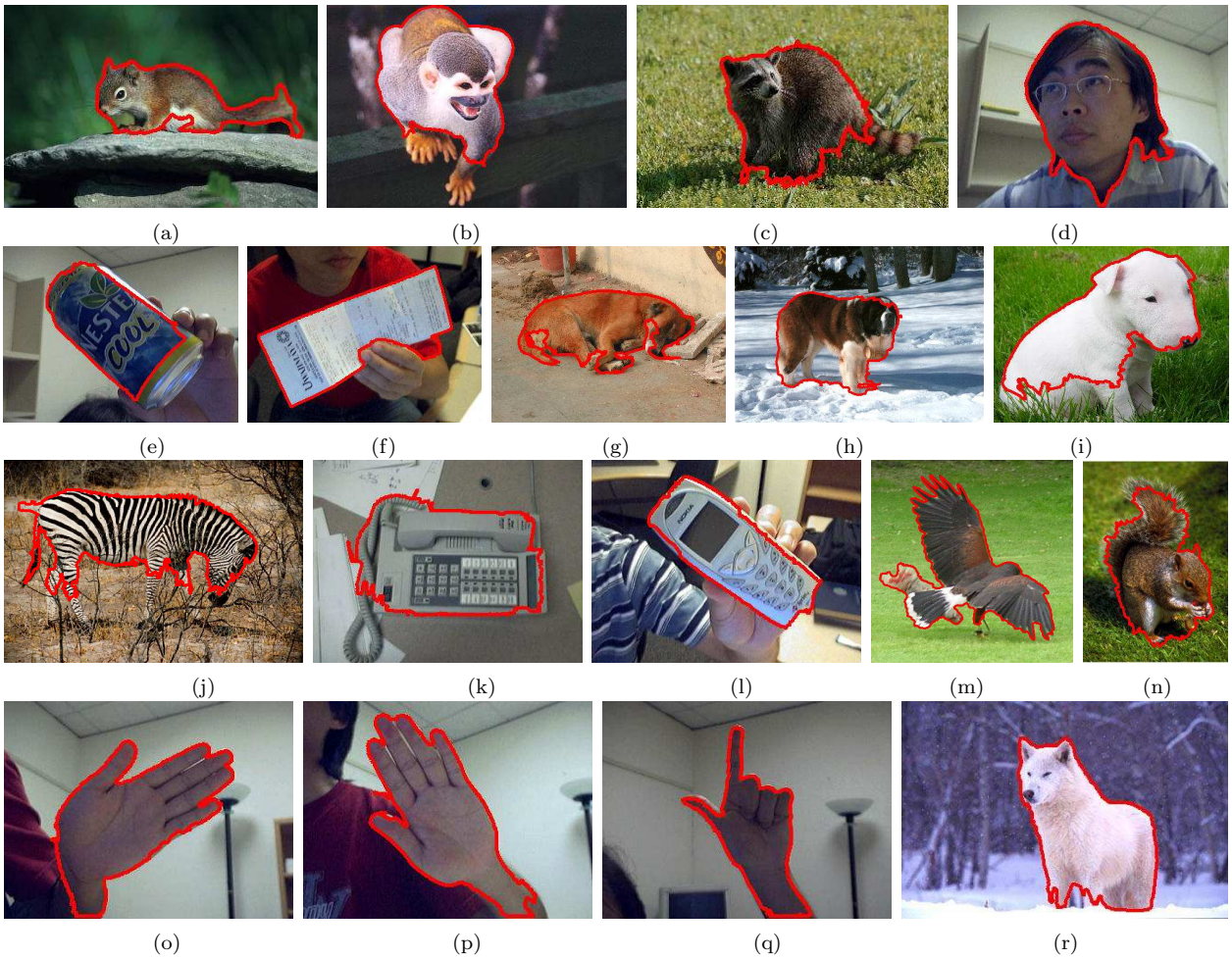


Figure 12: Segmentation results of other general objects.

[9] Yuri Boykov, Vivian S. Lee, Henry Rusinek, and Ravi 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*, volume LNCS 2208, pages 1058–1066, 2001.

[10] Yuri Boykov, Olga Veksler, and Ramin Zabih. Efficient approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239, November 2001.

[11] Vicent Casselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[12] Tony F. Chan and Luminita A. Vese. Active contours without edges. *IEEE Transaction on Image Processing*, 10(2):266–277, Feburary 2001.

[13] Laurent D. Cohen and Isaac Cohen. Finite-element methods for active contour models and balloons for 2-d and 3-d images. *IEEE Transaction on Pattern Analysis and Machine*

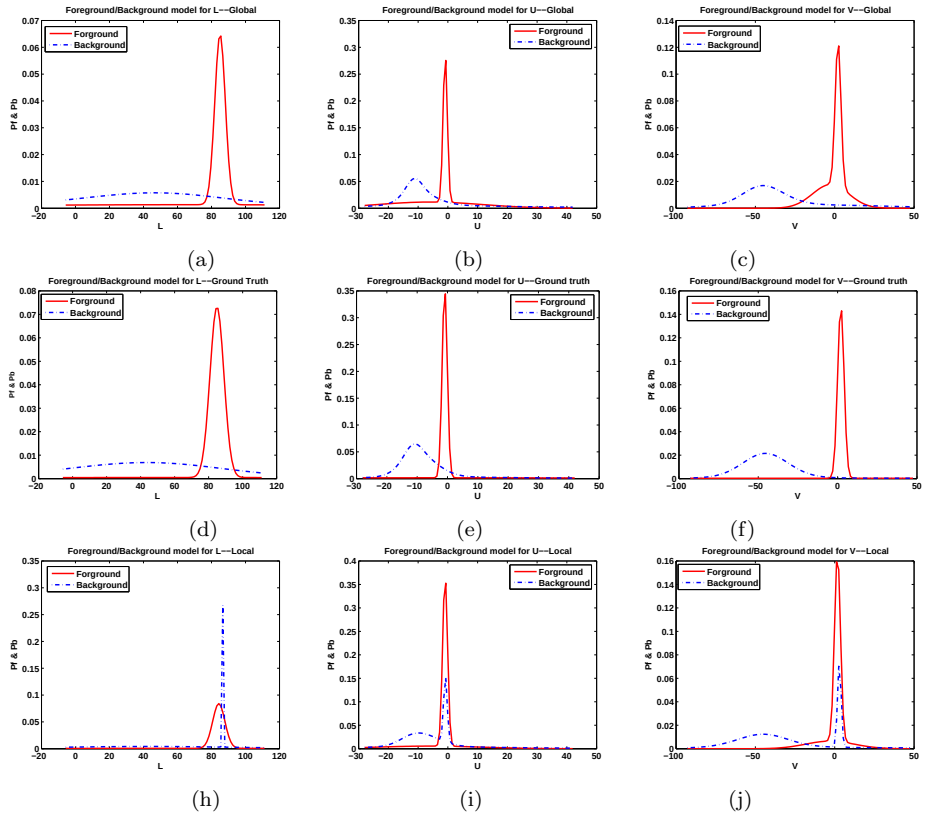


Figure 13: The estimated marginal distribution of foreground (red line) and background (blue line) on L (first column), U (second column) and V (last column). The first row presents these distributions obtained by our algorithm upon convergence. The second row presents the ground-truth distributions 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).

Intelligence, 15(11):1131–1147, 11 1993.

[14] Dorin Comaniciu and Peter Meer. Mean-shift: A robust approach toward feature space analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(5):1–18, May 2002.

[15] Adrian Corduneanu and Tommi Jaakkola. Continuation methods for mixing heterogenous sources. In *Proc. of Uncertaity in Artificial Intelligence*, pages 111–118, 2002.

[16] 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*, 39(1):1–38, 1977.

[17] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741, June 1984.

[18] Stéphanie Jehan-Besson, Michel Barlaud, and Gilles Aubert. Video object segmentation using eulerian region-based active contours. In *Proc. of IEEE International Conference on Computer Vision*, volume 1, pages 353–361, Vancouver, Canada, July 2001.

[19] Stéphanie Jehan-Besson, Michel Barlaud, and Gilles Aubert. Shape gradients for histogram segmentation using active contours. In *Proc. of IEEE International Conference on Computer Vision*, volume 1, pages 408–415, Nice,Côte d'Azur,France, Octobor 2003.

[20] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1987.

[21] Junmo Kim, John W. Fisher III, Anthony J. Yezzi, Müjdat Çetin, and Alan S. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Transaction on Image Processing*, 14(10):1486–1502, Octobor 2005.

[22] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Feburary 2004.

[23] Neil D. Lawrence and Bernhard Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In C. Brodley and A. P. Danyluk, editors, *Proc. of the International Conference in Machine Learning*, pages 306–313, San Francisco, CA, 2001. Morgan Kauffman.

[24] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D. Fox. Level set evolution without re-initialization: A new variational formulation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 430–436, San Diego, CA, June 2005.

[25] Ravikanth Malladi, James A. Sethian, and Baba C. Vemuri. Shape modeling with front propagation: a level set approach. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(2):158–175, Feburary 1995.

[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. In *8th IEEE International Conference on Computer Vision*, volume 2, pages 416–423, July 2001.

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

[28] Stanley Osher and James A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulation. *Journal of Computational Physics*, 79:12–49, 1988.

[29] Nikos Paragios and Rachid Deriche. Geodesic active contours for supervised texture segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1034–1040, 1999.

[30] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, pages 223–247, 2002.

[31] Nikos Paragios, Olivier mellina Gottardo, and Visvanathan Ramesh. Gradient vector flow fast geodesic active contours. In *Proc. of IEEE International Conference on Computer Vision*, pages 67–73, Vancouver, Canada, July 2001.

[32] Danping Peng, Barry Merriman, Stanley Osher, Hongkai Zhao, and Myungjoo Kang. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulation. *Journal of Computational Physics*, 155:410–438, 1999.

[33] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut"– interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH'04)*, pages 309–314, 2004.

[34] Mikaël Rousson, Thomas Brox, and Rachid Deriche. Active unsupervised texture segmentation on a diffusion based feature space. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 699–704, Madison, Wisconsin, June 2003.

[35] Christophe Samson, Laure Blanc-Féraud, Gilles Aubert, and Josiane Zerubia. A level set model for image classification. *International Journal of Computer Vision*, 40(3):187–197, March 2000.

[36] Yonggang Shi and William Clement Karl. Real-time tracking using level sets. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 34–41, San Diego, CA, June 2005.

[37] Andy Tsai, Jr. Anthony Yezzi, and Alan S. Willsky. A curve evolution approach to smoothing and segmentation using the mumford-shah functional. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1124, Hilton Head Island, South Carolina, June 2000.

[38] Zhuowen Tu. An integrated framework for image segmentation and perceptual organization. In *Proc. of IEEE International Conference on Computer Vision*, Beijing, China, Octobor 2005.

[39] Olga Veksler. *Efficient Graph-Based Energy Minimization Methods in Computer Vision.* PhD thesis, Cornell University, 1999.

[40] Zhengyou Zhang and Liwei He. Notetaking with a camera: Whiteboard scanning and image enhancement. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 533–536, Montreal, Quebec, Canada, May 2004.

[41] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Transaction on Pattern Recognition and Machine Intelligence*, 18(9):884–900, 9 1996.

[42] Xiaojin Zhu, Jie Yang, and Alex Waibel. Segmenting hands of arbitrary color. In *Proc. IEEE International Conference on Automatic Face Recognition*, pages 446–453, Grenoble, France, March 2000. IEEE Computer Society.