Tri-Map Self-Validation Based on Least Gibbs Energy for Foreground Segmentation

Xiaomeng Wu
wu.xiaomeng@lab.ntt.co.jp
Kunio Kashino
kashino.kunio@lab.ntt.co.jp

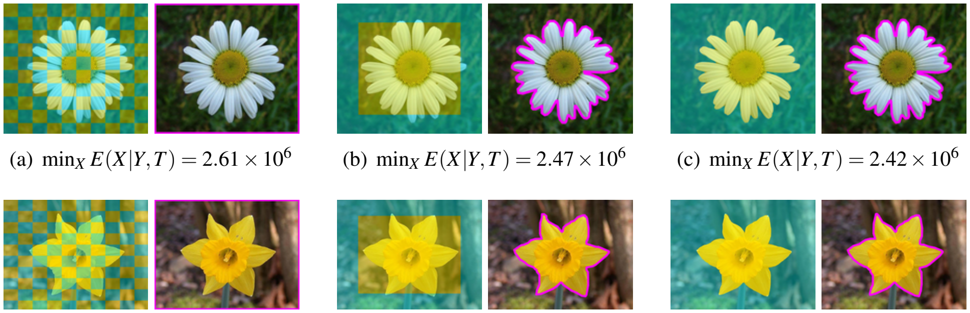
NTT Communication Science Laboratories
3-1, Morinosato Wakamiya Atsugi-shi
Kanagawa, Japan 243-0198

Abstract

The Bayesian framework forms a solid foundation for image segmentation. With this as a basis, an image is modeled as a Markov random field (MRF) with observations incorporated with a given tri-map. Although MRF-based methods have proved successful in interactive or supervised foreground segmentation, high-quality segmentation can be obtained only when the tri-map is sufficiently discriminative. We argue that the least Gibbs energy can be formulated as a goal function of a tri-map and can be a powerful means of validating the separability of predefined feature distributions. Further, we propose a split-and-validate strategy for decomposing the complex problem into a series of tractable subproblems, and suboptimal tri-map optimization is gradually achieved by making decisions between cluster-level operations. The splitting is determined by a novel combination of Bregman hierarchical clustering and an information theoretic method for realizing non-parametric clustering. We have evaluated our method against the Oxford Flower 17 and Caltech-UCSD Bird 200 benchmarks and show the superiority of tri-map self-validation in unsupervised foreground segmentation tasks.

1 Introduction

As an intermediate process, foreground segmentation plays an important role in high-level vision tasks, *e.g*. image matting and image classification. Of previously reported research, a large percentage is made up of Markov random field (MRF) based studies [3, 4, 5, 14, 22, 23], in which statistically optimal segmentation maximizes the posteriori probability given observations incorporated with a given tri-map. A tri-map is an indefinite assignment of each pixel in the source image to one of two or three classes: foreground, background, and/or unknown. The segmentation uses the information from the foreground and background regions to reclassify each pixel into foreground or background classes. MRF-based methods naturally represent the segmentation accuracy and spatial coherence within the Bayesian framework, but mainly under the assumption that a sufficiently discriminative tri-map is given, *e.g*. specified by user interaction [3, 4, 14, 23], learned from ground-truth segmentations [22], or weakly supervised by using class information [5, 22]. This assumption makes them unsuitable for generic applications, *e.g*. when a large number of images are to be processed [3, 4, 14, 23] or when the training data are unavailable [5, 22]. Although some attempts have been made to improve the discriminative power of feature descriptors [11, 14]



(a) $\min_X E(X|Y,T) = 2.61 \times 10^6$

(b) $\min_X E(X|Y,T) = 2.47 \times 10^6$

(c) $\min_X E(X|Y,T) = 2.42 \times 10^6$

(d) $\min_X E(X|Y,T) = 3.47 \times 10^6$

(e) $\min_X E(X|Y,T) = 3.34 \times 10^6$

(f) $\min_X E(X|Y,T) = 3.20 \times 10^6$

Figure 1: Different tri-maps T exhibit differences in their least Gibbs energies (LGE) $\min_X E(X|Y,T)$ when incorporated in the segmentation of the same image Y. Consider the tri-maps on the left of each sub-figure. The images on the right show the segmentation X that obtains the minimal cut of the MRF graph. A less ambiguous tri-map is usually conducive to a lower LGE than a less discriminative one, even if they lead to similar segmentations. Based on this observation, in this paper, we introduce the LGE as a measure that captures the separability of predefined appearance distributions and incorporate it in a split-and-validate strategy for foreground segmentation.

or the MRF model [6, 26] given a low-quality tri-map, very little attention has been paid to enhancing the discernment of the tri-map itself. This constitutes the main problem that we tackle in this paper.

In contrast to the previous studies, which depended on strong assumptions, our aim is unsupervised foreground segmentation under only one weak (realistic) assumption. Namely, we assume that the location of a foreground object is a normal deviate in the image space, whose expectation lies near the center of the image, and with a sufficiently low standard deviation. Based on this assumption, a simple solution is to assume a rectangle of sufficient scale centered in the image as the tri-map [6], as shown in Fig. 1(b) and 1(e). The segmentation can thus be obtained by minimizing the configuration energy combining boundary regularization with appearance models predefined on the basis of the tri-map. When the distributions offer very low separability, as shown in Fig. 1(a) and 1(d), the appearance models become non-contributory and the minimization over-fits the boundary regularization. As a consequence, we expect to see a high minimal energy. When two tri-maps lead to the same segmentation, i.e. to the same boundary regularization, as shown in Fig. 1(b) and 1(c), the tri-map with the larger overlap in the feature distributions indicates a higher entropy.

Using this observation as a basis, we develop our contributions as follows. We formulate the least Gibbs energy as a tri-map goal function and propose a split-and-validate strategy to decompose the complex problem into a series of tractable subproblems (Section 2). A suboptimal tri-map optimization is gradually obtained by making decisions between cluster-level operations. The splitting is determined by a novel combination of Bregman hierarchical clustering and an information theoretic method to realize non-parametric clustering and to balance the clustering quality and computational efficiency (Section 3). We comprehensively evaluate our method using the Oxford Flower 17 and Caltech-UCSD Bird 200 benchmarks, and present extensive experiments demonstrating the superiority of self-validation in unsupervised foreground segmentation tasks (Section 4).

1.1 Literature Review

Energy-based segmentation methods [8, 28] wherein segment characteristics are modeled through MRFs have interested researchers over the past several years. A desired segmentation is defined as one that maximizes the product of the class conditional distribution (characterizing features such as intensity) and the a priori probability distribution (imposing spatial connectivity constraints). Among MRF-derived methods, there is a prominent category of techniques [3, 4, 7, 9, 14, 23, 26] that employ graph representations for image segmentation. Boykov and Jolly [3] represented an image as an undirected graph and used a configuration energy that incorporates background-foreground appearance models derived from intensity histograms of a tri-map, and boundary regularization as soft constraints. A graph-cut framework based on the min-cut criterion [4] was utilized to uncover a globally optimal solution as the final segmentation outcome. This work was further enhanced by Rother *et al.* [23], who devised an iterated graph-cut methodology called *GrabCut*, where color information is incorporated into the configuration energy using a Gaussian mixture model. Han *et al.* [14] established a color image segmentation algorithm by extending Rother *et al.*'s methodology to accommodate multi-scale nonlinear structure tensor texture features. Feng *et al.* [9] proposed an unsupervised extension of the binary graph cut known as a graduated graph cut, with an architecture that is capable of the self-validated labeling of MRFs. Apart from the above procedures, methods involving label costs [7] and segmentation overlap costs [26] in the configuration energy have been developed for driving various imaging applications.

In the literature related to foreground segmentation, a significant effort has been devoted to the development of supervised techniques for achieving results that are more tailored towards user requirements. These techniques can also be categorized in terms of interactive methods [3, 4, 14, 23] and learning-based methods [1, 5, 16, 19, 20, 22]. Interactive methods start with user-specified contours [3, 4] or regions [14, 23] to compute the foreground-background appearance model. Degrees of interactive effort range from editing individual pixels, at the labor-intensive extreme, to touching an image in a few locations. These methods are of great practical importance in image editing, but are labor-consuming when a large number of images are to be processed. Nilsback and Zisserman [22] assumed that ground-truth segmentations are available and the objects are star-shaped polygons, and they described an algorithm for segmenting flowers in color images. Najjar and Zagrouba [20] as well as Angelova *et al.* [1] removed the assumption of the geometric shape and proposed segmentation methods for flower classification. Another group of studies, known as co-segmentation studies, considered sets of images where the foreground appearance share similarities that can be leveraged to obtain accurate segmentations. Discriminative learning on super-pixels is used by Joulin *et al.* [16] for simultaneously enforcing spatial smoothness as well as finding the foreground-background boundary in a super-pixel (feature) space. Unlike this study, Chai *et al.* [5] decouples spatial smoothness enforcement and the classification of super-pixels, so that these steps are performed consecutively rather than simultaneously. Co-segmentation ranks as one of the most advanced research topics in the literature. However, this group of methods suffers from limited utility when the ground-truth class information is unavailable beforehand.

2 Tri-Map Self-Validation

2.1 Least Gibbs Energy

In terms of MRFs, the statistically optimal segmentation $\hat{X}$ maximizes the a posteriori probability pertaining to an observed image Y and a tri-map T, which is equivalent to minimizing the Gibbs energy $E(X|Y,T)$:

$$E(X|Y,T) = \sum_{p}\sum_{\alpha} U_p^{(\alpha)}(y_p|T)\delta(\alpha,x_p) + \sum_{p,q} \frac{1-\delta(x_p,x_q)}{\|p-q\|}\exp(-\beta\|y_p-y_q\|) \tag{1}$$

where the right terms are known as the likelihood (first) and coherence (second) energy functions at the pixel level. The likelihood energy represents the goodness of labeling pixel p by x_p. Here, $x_p \in \{0,1\}$ and y_p, respectively, are the label and the descriptor of each pixel p. $\alpha \in \{0,1\}$ is the label denoting the foreground or background and $\delta(\cdot,\cdot)$ denotes the Kronecker delta. $U_p^{(\alpha)}$ represents the appearance model, *e.g*. the negative log likelihood of a histogram [1, 3, 22], a Gaussian mixture model (GMM) [14, 23], or a model learned from a texon [18] or *k*-means centroids [9]. The coherence energy extends the Potts model, and measures the boundary regularization feasibility, which represents the spatial connectivity.

Several example tri-maps are shown in Fig. 1. Here, we define the least Gibbs energy (LGE) as follows:

$$LGE(T|Y) = \min_{X} E(X|Y,T) \tag{2}$$

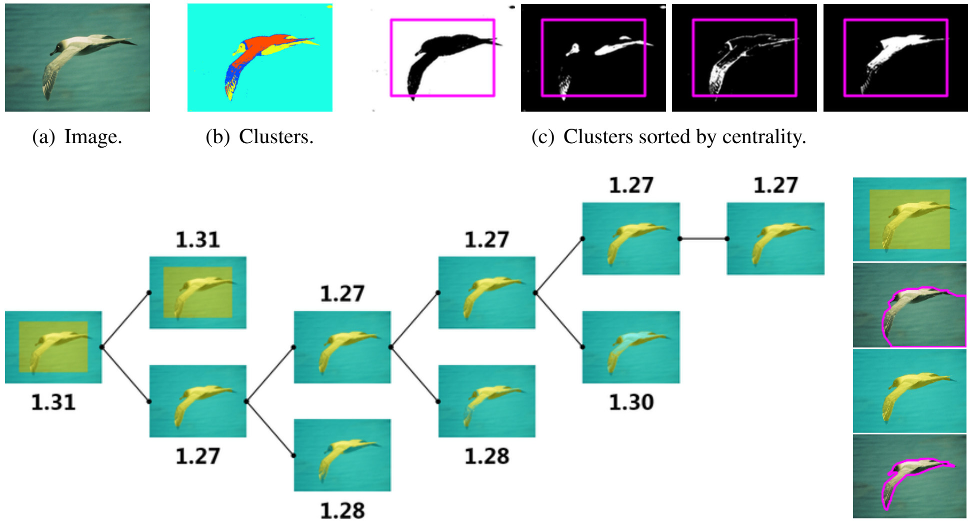
Obviously, LGE is a function of the variable T with a given observation Y, and is no longer dependent on the segmentation X. As shown in Fig. 1, a less ambiguous tri-map T usually leads to a lower $LGE(T|Y)$ than a less discriminative one, even if they lead to similar segmentations.

This observation is also supported by the convergence property of iterated graph cut [23]. This method considers the segmentation $\hat{X}$ minimizing the Gibbs energy in the previous iteration as the tri-map $T^{(t)}$ of the current iteration, and optimizes energy function $E(X|Y,T^{(t)})$ over the segmentation. These steps are repeated recursively. In practice, $\min_X E(X|Y,T^{(t)})$ decreases monotonically and the iteration is guaranteed to converge to a local minimum sensitive to the initial tri-map $T^{(0)}$. If we assume the fixed point $\hat{X}$ after the convergence corresponds to a *perfect* segmentation, the tri-map $\hat{T}$ that led to this *perfect* segmentation can be considered the most desired tri-map and that fits the data distribution in the foreground and background. The convergence of the iterated graph cut indicates that the $LGE(\hat{T}|Y)$ of $\hat{T}$ must be lower than that of any other arbitrary tri-map. On the basis of this observation, we develop our split-and-validate method in the next section.

2.2 Split and Validate

A desired tri-map $\hat{T}$ can be defined as one that minimizes $LGE(T|Y)$, more specifically $\hat{T} = \arg\min_T \min_X E(X|Y,T)$. Straightforwardly solving this problem is NP-hard. Instead, we propose a split-and-validate strategy, which decomposes the complex problem into a series of tractable subproblems. The splitting is determined by the non-parametric clustering method described in Section 3. After splitting, the image is abstracted as a set of pixel clusters. Our tri-map refinement is based on the following two types of cluster-level operations:

1 (Retaining) *For a tri-map T, keeping T unchanged, as denoted by $T \leftarrow T$.*



(a) Image. (b) Clusters. (c) Clusters sorted by centrality.

(d) Refinement by cluster-level operations. (e) Results.

Figure 2: Working flow of tri-map self-validation. Validation is initialized with the rectangle in the center (50% of the image size). In 2(c), the pixels in the current cluster are in white. In each *column* of 2(d), the upper image is the current T while the lower image is the tentative T' to be validated. The reals above/below each image are the LGEs (unit: 10^6) of T or T'. The segmentations derived from $T^{(0)}$ and the refined $\hat{T}$ are compared in 2(e).

2 (Contracting) *For a tri-map $T = \{T_B, T_F\}$, in which T_B and T_F are background and foreground regions, and a pixel cluster c, subtracting c from T_F and adding c to T_B, as denoted by $T \leftarrow \{T'_B, T'_F\}$ with $T'_B = T_B \cup c$ and $T'_F = T_F \setminus c$.*

The self-validation scheme is discretized to a tree-structured evolution process (Fig. 2(d)). The tri-map $T^{(0)} = \{T_B^{(0)}, T_F^{(0)}\}$ is first treated as a rectangle in the center, i.e. we assume the location of a foreground object to be a normal deviate in the image space whose expectation lies near the center. $T^{(0)}$ can be modeled as a binary MRF. Using the appearance model $U_p^{(\alpha)}$ and the corresponding energy assignment (Eq. 1), we can obtain $LGE(T^{(0)}|Y)$.

All pixel clusters $c_1, c_2, \cdots, c_G$ are then sorted in ascending order of centrality, which is defined as the ratio of $\|T_F^{(0)} \cap c\|$ over $\|T_F^{(0)}\|$. This is motivated again by the above assumption, and a cluster of pixels is considered more likely to belong to the foreground if its location is closer to the center of the image. $T^{(0)}$ is then arguably refined by the operation *contracting* with the cluster c at the top of the sorted queue, which leads to a tentative tri-map $T'^{(0)}$ and in consequence $LGE(T'^{(0)}|Y)$. A tri-map T is contract-able if the operation *contracting* leads to a lower LGE than *retaining*, i.e. $LGE(T'|Y) < LGE(T|Y)$. If so, we update T to T', and continue this process iteratively until all clusters are incorporated in the validation. Then, we obtain the final segmentation by using an iterated graph cut [23] with the refined $\hat{T}$. As shown in Fig. 2(e), for the tri-map initialized without supervision, the overlap between the foreground in yellow and the background in blue is too strong, which makes the image hard to segment. However, self-validation is able to define the desired tri-map accurately and in consequence define the segment automatically.

3 Distortion-Based Bregman Hierarchical Clustering

3.1 Determining Number of Pixel Clusters

A clear problem with pixel clustering is finding a way to specify the number of clusters to detect. Under-clustering may lead to significant confusion between foreground and background pixels, while over-clustering could result in a larger number of validations. Of the algorithms that solve this problem [12, 13, 15, 17, 24], we chose the distortion-guided method [24] because it is non-parametric, mathematically supported, and computationally simplex.

Consider a p-dimensional variable Y consisting of a mixture of G distributions. Given a set of k cluster centroids, we can define the pooled mean Mahalanobis distance per dimension between Y and the centroids as a *distortion* d_k of the k-clustering. Sugar and James [24] proved that the transformed distortion $d_k^{-p/2}$ is approximately zero for $k < G$, and then jumps suddenly and begins increasing linearly for $k \geq G$. Jumps in $d_k^{-p/2}$ then signify reasonable choices for k. Sugar and James [24] generated the k cluster centroids using a k-means algorithm with a varying $k \in [1, K]$. This requires K-fold computations for k-means as well as for distortion calculation. Instead, we propose using a Bregman hierarchical clustering (BHC) method to generate clusters.

3.2 Bregman Hierarchical Clustering

Garcia and Nielsen [10] have shown that agglomerative hierarchical clustering can be approximated and accelerated by considering the data to be a mixture f of $N \geq G$ distributions, *e.g*. Gaussian or Poisson distributions etc., and instantiating the hierarchical clustering to this mixture with Bregman divergence. BHC [10] leverages the symmetric Bregman divergence between two distributions as the linkage criterion, which allows us to identify the two closest distributions in order to merge them into a meta-distribution. As a result of this merging, the number of distributions (starting at N) decreases by one after each iteration until there is one distribution containing all the components.

BHC determines the optimal clustering g as the one with the minimum number of components that reaches a minimum approximation quality $d(f,g) \leq \tau$ defined by the user, where $d(\cdot,\cdot)$ is the symmetric Bregman divergence between the mixtures. Instead, we adapt the distortion-guided criterion [24] to BHC to achieve a non-parametric determination of g. Here, the pooled mean Mahalanobis distance described in Section 3.1 can be reduced to the pooled mean variances of Y in each cluster, which equals $\mathrm{tr}(\Sigma)$ with Σ denoting the covariance matrix if we consider a multivariate Gaussian distribution. The determination thus requires only one computation, namely the initialization of the mixture of N Gaussian distributions, following which the natural parameters $\{\mu, \Sigma, \omega\}$ of each distribution are calculated. In the merging phase, $\hat{\Sigma}$ of the meta-distribution $\hat{c}$ merged from the previous distributions c_i and c_j in each iteration can be efficiently updated (not requiring N-fold computations for generating clusters or for distortion calculation). We compare the non-parametric solution of clustering generation with the original BHC in Section 4.3.

Table 1: Comparison of various self-validation criteria [1].

MJI (%)				**MNHS (%)**			
GC [23]	**LGE**	**LLE**	**JD**	**GC [23]**	**LGE**	**LLE**	**JD**
89.3	**91.7**	90.8	57.3	95.9	**96.7**	96.4	84.9

[1] OF17 is used in this experiment. Self-validation is initialized with the rectangle in the center (50% of the image size). Given a tri-map, the segmentation is achieved by a five-iteration graph cut for all methods. MJI: Mean Jaccard Index. MNHS: Mean Normalized Hamming Similarity. GC: Graph Cut. LGE: Least Gibbs Energy. LLE: Least Likelihood Energy. JD: Jaccard Distance.

4 Experimentation

4.1 Setting

For our evaluation, we use the Oxford Flower 17 (OF17) [21] and Caltech-UCSD Birds 200 (CB200) [27] datasets. OF17 contains 17 flower species with 80 images per category. 846 of them have hand-annotated segmentations. CB200 contains 200 bird categories and 6033 images in total, in which rough segmentation masks are provided and this allows us to compare segmentation accuracies. The foreground-background overlap in terms of color distributions is very strong in CB200 making this dataset much more challenging than OF17.

Given a ground-truth segmentation, the accuracy can be expressed in two ways. Let TP, FP, FN, and TN denote the numbers of true-positive, false-positive, false-negative, and true-negative pixels, respectively. One criterion [5, 16, 22] is the mean of the Jaccard index (MJI) between the estimated foreground and the ground-truth foreground, in which JI equals $TP/(TP+FP+FN)$. The other criterion [5, 16] is the mean of the normalized Hamming *similarity* (MNHS), in which the NHS exactly equals the accuracy defined in binary classification. As suggested by Garcia and Nielsen [10], we chose $N = 32$ as the number of Gaussian components used in clustering initialization (Section 3.2).

4.2 Comparison of Self-Validation Criteria

To the best of our knowledge, few studies have tackled the formulation of the ambiguity of a tri-map. In this section, we compare the LGE defined in Eq. 2 with two alternatives. It should be noted that the coherence energy based on the Potts model in Eq. 1 is completely independent from the tri-map T. We first determine how much effect this independent term has on tri-map self-validation by defining the criterion as $E_L(Y|\hat{X},T)$, which is the least likelihood energy (LLE) at the image level. On the other hand, from the convergence property of iterated graph cut, we can deduce that a desired tri-map should be the desired segmentation itself while an ambiguous tri-map should have limited shape similarity to its corresponding segmentation. The criterion can thus be defined by a shape similarity, namely the Jaccard distance (JD) between T and its corresponding segmentation $\hat{X}$. Table 1 compares the segmentation accuracy obtained using various criteria.

From Table 1, we can see that both LGE and LLE outperformed the initial tri-map, while LGE slightly outperformed LLE. This is because, although T is independent from the Potts model, the minimization of the Gibbs energy does depend on this coherence energy. A desired tri-map should enable us to realize a segmentation with both a small feature distribution overlap and a low spatial connectivity between the segments, so LLE alone is insufficiently informative to capture the discernment of T. JD failed to correctly measure the tri-map dis-

Table 2: Comparison of non-parametric BHC and previous studies [1].

OF17	**MJI (%)**					**MNHS (%)**				
Tri-Map Scale	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
GC [23]	89.3	87.6	85.7	81.8	76.5	95.9	94.5	93.1	90.1	85.0
BHC ($k = 9$) [10]	90.9	89.7	86.2	81.9	70.0	96.5	95.7	93.5	89.6	78.3
BHC-BD ($\tau = 7.0$) [10]	**91.8**	90.2	**89.0**	84.9	79.4	**96.8**	95.9	94.8	91.4	86.1
BHC-RD	91.7	**91.0**	88.7	**86.0**	**80.4**	**96.8**	**96.3**	**94.9**	**92.5**	**86.8**

CB200	**MJI (%)**					**MNHS (%)**				
Tri-Map Scale	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
GC [23]	35.7	32.5	28.9	25.5	22.7	68.2	60.6	52.3	44.5	37.5
BHC ($k = 9$) [10]	36.4	33.1	28.9	25.2	21.7	71.8	65.0	56.8	48.9	40.4
BHC-BD ($\tau = 4.0$) [10]	**36.6**	**33.9**	**30.9**	27.9	24.6	72.7	66.7	60.5	54.2	46.8
BHC-RD	36.0	33.5	**30.9**	**28.6**	**25.8**	**72.9**	**67.1**	**61.2**	**56.1**	**49.5**

[1] For BHC and BHC-BD, the parameters in the brackets are those resulting in the corresponding highest MJI. OF17: Oxford Flower 17. CB200: Caltech-UCSD Birds 200. BHC: Bregman Hierarchical Clustering. BHC-BD: BHC using Bregman Divergence as stopping criterion. BHC-RD: BHC using Rate Distortion as stopping criterion.

cernment. This may be because, although the shape of a desired tri-map should be similar to that of its segmentation, an ambiguous tri-map does not necessarily reflect this assumption.

4.3 Comparison of Clustering Criteria

In this section, we compare the non-parametric BHC with the BHC methods that use various stop criteria. We use BHC to denote the clustering constraining the number $k \in (1, 10]$ of components, and use BHC-BD to denote the clustering constraining the minimum approximation quality $\tau \in [0.0, 10.0]$. Although the tri-map can be centrally initialized without any supervision, it is hard to finely determine the scale of the tri-map without any knowledge of the scale of the object of interest. A generic method should enable reliable segmentations over various configurations of tri-map initialization. To evaluate the methods from this viewpoint, we vary the rectangular scale from 50% to 90% of the image size. Table 2 compares the best performance of various methods.

In general, all self-validation methods outperformed GC [23]. The superiority of BHC-BD to BHC demonstrates that constraining the quality of distributions is more appropriate than constraining the number of distributions. Even so, BHC-BD requires tuning to achieve its best performance shown in Table 2. In unsupervised foreground segmentation, cross validation is obviously impossible since the training data are unavailable, which limits the utility of BHC-BD. In contrast, our method is non-parametric, i.e. it is not dependent on the specification of the quality or the number of distributions. It achieved the highest performance in most cases, and even for the few exceptions, was comparable to the finely-tuned BHC-BD. Figure 3 compares the segmentations initialized by the same tri-map (50% of the image size). The segmentation without tri-map optimization [23] failed in most cases because of the small object size and the large number of *background* pixels mislabeled as *foreground* (example first from the right). In contrast, our method successfully removed confusing regions even when the foreground showed high color similarity to the background (example second from the right). The higher robustness of our method as regards low-quality tri-map initialization

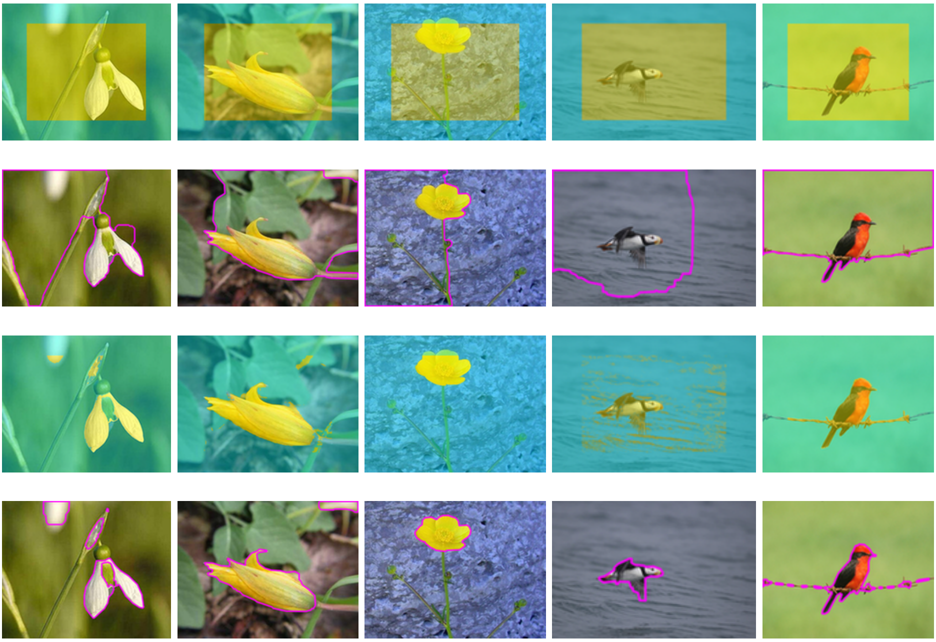


Figure 3: Example of tri-map optimization and segmentation. From top to bottom: initialized tri-map (used in both GC [23] and our method), segmentation of GC, optimized tri-map, and our segmentation.

is also reflected in Table 2. For the challenging CB200 dataset, our method achieved a 3.1% MJI improvement and a 12.0% MNHS improvement compared with our baseline when the rectangular scale was enlarged to 90%. Many more comparative results are provided in the supporting documentation.

4.4 Comparison with State-of-Art Methodologies

In this section, we compare our method with advanced studies, including both supervised [5, 16, 20, 22] and unsupervised [2, 25] methodologies, by using the OF17 dataset. Because both coarse-grained (flower) and fine-grained class information is available in this dataset, it is inevitable that methods interleaving the segmentation and construction of class-specific appearance models have an advantage over any unsupervised method including ours. Nevertheless, it is interesting to see how well self-validation can fare against such methods. The comparison is shown in Table 3. In the experiments, self-validation performed worse than the flower-geared method [22] and Chai *et al.*'s method [5], where the latter uses both coarse- and fine-grained class information. Apart from that, our method is competitive, and in particular it outperforms Joulin's [16] and Najjar's [20] co-segmentation methods. Also, our method is the top-performing unsupervised method.

Table 3: Segmentation performance on OF17 dataset reported in the literature.

Supervised Method	**MJI (%)**	**MNHS (%)**
Nilsback and Zisserman [22]	93.0	–
Joulin *et al.* [16]	75.8	86.6
Chai *et al.* [5]	**94.7**	**98.3**
Najjar and Zagrouba [20]	84.0	–
Unsupervised Method	**MJI (%)**	**MNHS (%)**
Aydin and Ugur [2]	87.0	–
Suta *et al.* [25]	90.0	89.0
Our Method [1]	**91.7**	**96.8**

[1] Note that our method differ completely from Chai *et al.*'s method [5]. The former is not a co-segmentation method, and does not require any class information.

5 Conclusion

We presented a tri-map self-validation method that is arguably simple, and yet it outperformed previously reported techniques based on unsupervised methods [2, 25] and was competitive with supervised methods [5, 16, 20, 22], which treat images dependently. We have shown that the least Gibbs energy can be a strong cue for capturing the discriminative power of a tri-map, which is further incorporated in a split-and-validate strategy for non-parametric tri-map optimization. In the future, we intend to examine whether we can optimize the tri-map and infer the appearance models simultaneously such that a simpler and more efficient optimization can be easily incorporated into any graph cut-based segmentation application. On the other hand, we have not discussed the loss in the translation to the split-and-validate problem as an approximation of the NP-hard problem mentioned in Section 2.2. These issues constitute our future direction.

References

[1] Anelia Angelova, Shenghuo Zhu, and Yuanqing Lin. Image segmentation for large-scale subcategory flower recognition. In *WACV*, pages 39–45, 2013.

[2] Dogan Aydin and Aybars Ugur. Extraction of flower regions in color images using ant colony optimization. *Procedia CS*, 3:530–536, 2011.

[3] Yuri Boykov and Marie-Pierre Jolly. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. In *ICCV*, pages 105–112, 2001.

[4] Yuri Boykov and Vladimir Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

[5] Yuning Chai, Victor S. Lempitsky, and Andrew Zisserman. BiCoS: A Bi-level co-segmentation method for image classification. In *ICCV*, pages 2579–2586, 2011.

[6] Daniel Chen, Brenden Chen, George Mamic, Clinton Fookes, and Sridha Sridharan. Improved GrabCut Segmentation via GMM Optimisation. In *DICTA*, pages 39–45, 2008.

[7] Andrew Delong, Anton Osokin, Hossam N. Isack, and Yuri Boykov. Fast Approximate Energy Minimization with Label Costs. *International Journal of Computer Vision*, 96 (1):1–27, 2012.

[8] Aristeidis Diplaros, Nikos A. Vlassis, and Theo Gevers. A Spatially Constrained Generative Model and an EM Algorithm for Image Segmentation. *IEEE Transactions on Neural Networks*, 18(3):798–808, 2007.

[9] Wei Feng, Jiaya Jia, and Zhi-Qiang Liu. Self-Validated Labeling of Markov Random Fields for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(10):1871–1887, 2010.

[10] Vincent Garcia and Frank Nielsen. Simplification and hierarchical representations of mixtures of exponential families. *Signal Processing*, 90(12):3197–3212, 2010.

[11] L. Gavish, Lior Shapira, Lior Wolf, and Daniel Cohen-Or. One-sided object cutout using principal-channels. In *CAD/Graphics*, pages 448–451, 2009.

[12] Cyril Goutte, Peter Toft, Egill Rostrup, Finn A. Nielsen, and Lars Kai Hansen. On Clustering fMRI Time Series. *NeuroImage*, 9(3):298 – 310, 1999.

[13] Cyril Goutte, Lars Kai Hansen, Matthew G. Liptrot, and Egill Rostrup. Feature-space clustering for fMRI meta-analysis. *Human Brain Mapping*, 13(3):165–183, 2001.

[14] Shoudong Han, Wenbing Tao, Desheng Wang, Xue-Cheng Tai, and Xianglin Wu. Image Segmentation Based on GrabCut Framework Integrating Multiscale Nonlinear Structure Tensor. *IEEE Transactions on Image Processing*, 18(10):2289–2302, 2009.

[15] Mehrdad Honarkhah and Jef Caers. Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling. *Mathematical Geosciences*, 42(5):487–517, 2010.

[16] Armand Joulin, Francis R. Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010.

[17] R. Lleti, M.C. Ortiz, L.A. Sarabia, and M.S. Sanchez. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1):87 – 100, 2004.

[18] Jitendra Malik, Serge Belongie, Thomas K. Leung, and Jianbo Shi. Contour and Texture Analysis for Image Segmentation. *International Journal of Computer Vision*, 43 (1):7–27, 2001.

[19] Fanman Meng, Hongliang Li, King Ngi Ngan, Liaoyuan Zeng, and Qingbo Wu. Feature Adaptive Co-Segmentation by Complexity Awareness. *IEEE Transactions on Image Processing*, 22(12):4809–4824, 2013.

[20] A. Najjar and E. Zagrouba. Flower image segmentation based on color analysis and a supervised evaluation. In *Communications and Information Technology (ICCIT), 2012 International Conference on*, pages 397–401, 2012.

[21] Maria-Elena Nilsback and Andrew Zisserman. A Visual Vocabulary for Flower Classification. In *CVPR (2)*, pages 1447–1454, 2006.

[22] Maria-Elena Nilsback and Andrew Zisserman. Delving deeper into the whorl of flower segmentation. *Image Vision Comput.*, 28(6):1049–1062, 2010.

[23] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[24] Catherine A. Sugar and Gareth M. James. Finding the number of clusters in a dataset: An Information-Theoretic approach. *Journal of the American Statistical Association*, 98(463):750–763, 2003.

[25] L. Suta, F. Bessy, C. Veja, and M.-F. Vaida. Active contours: Application to plant recognition. In *Intelligent Computer Communication and Processing (ICCP), 2012 IEEE International Conference on*, pages 181–187, 2012.

[26] Meng Tang, Lena Gorelick, Olga Veksler, and Yuri Boykov. GrabCut in One Cut. In *ICCV*, pages 1769–1776, 2013.

[27] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[28] Lei Zhang and Qiang Ji. Image segmentation with a unified graphical model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(8):1406–1425, 2010.