LSI: Latent Semantic Inference for Natural Image Segmentation

Le Dong[a,*], Ning Feng[a], Qianni Zhang[b]

[a]*School of Computer Science and Engineering,University of Electronic Science and Technology of China (UESTC),2006 Xiyuan Avenue, Gaoxin West Zone, Chengdu, Sichuan, 611731, China*
[b]*School of Electronic Engineering and Computer Science, Queen Mary, University of London*

Abstract

We propose a novel label inference approach for segmenting natural images into perceptually meaningful regions. Each pixel is assigned a serial label indicating its category using a Markov Random Field(MRF) model. To this end, we introduce a framework for latent semantic inference of serial labels, called LSI, by integrating local pixel, global region, and scale information of an natural image into a MRF-inspired model. The key difference from traditional MRF based image segmentation methods is that we infer semantic segments in the label space instead of the pixel space. We first design a serial label formation algorithm named Color and Location Density Clustering (CLDC) to capture the local pixel information. Then we propose a label merging strategy to combine global cues of labels in the Cross-Region potential to grasp the contextual information within an image. In addition, to align with the structure of segmentation, a hierarchical label alignment mechanism is designed to formulate the Cross-Scale potential by utilizing the scale information to catch the hierarchy of image at different scales for final segmentation optimization. We evaluate the performance of the proposed approach on the Berkeley Segmentation Dataset and preferable results are achieved.

Keywords: Image Segmentation, MRF Model, Label Inference

*Corresponding author. Tel.:+86 13981763623; Fax: +86-28-61831655.
Email address: ledong@uestc.edu.cn (Le Dong)

1. Introduction

Image segmentation remains a highly challenging task in Computer Vision. By partitioning an image into a set of perceptually meaningful regions, it acts as an indispensable process for a range of middle level and high level vision tasks, such as object detection [1, 2, 3], object recognition [4, 5, 6], knowledge inference [9, 10, 13], and image understanding [11, 15].

Recent works have shown that employing color, texture, and any other contextual information lead to encouraging results for image segmentation. The link of all these information is usually obtained by learning various complex classifiers to control the segmentation results. Many significant segmentation algorithms [16, 17, 18, 20] generate and regard such link terms as preprocessed segments, and then try to utilize the advantages of superpixel-level segmentation method [20] which puts emphasis on finding the dissimilarity of pixels in the whole image. However, the superpixel-level segmentation method neglects the similarity of pixels in the local non-adjacent regions, which can also serve as another valuable element aside from dissimilarity of pixels that contributes to improve the segmentation performance. In this paper, different with traditional superpixel-level method, we are apt to exploit the characteristics of local pixels and global regions for enhancing image segmentation in a disciplined manner, which overcomes the limitation of homogeneous superpixel-based treatment. Although a unified framework called PISA [9] is proposed to generate image regions with the same purpose as ours, this method gives more attention to image saliency.

In this paper, we consider the segmentation task as a label inference problem. To solve the problem, we propose a new framework for latent semantic inference of serial label, called **LSI**. In this framework, we adapt a MRF-inspired model but focus on the crucial point of how to integrate the multiple information into the segmentation process. Unlike traditional methods based on MRF model [35, 33], the proposed $Cross - R\&S$ model can produce an estimate of the

number of object classes in the image by coupling potential functions defined on labels with the designed Cross-Region and Cross-Scale potentials.

In our LSI framework, the serial labels are obtained through a color and location density clustering (CLDC) algorithm in the label formation phase to catch the local characters of a natural image. Then, the next step is to assign each pixel a unique serial label with a MRF model by minimizing the energy function of labels. Although [12, 14] try to minimize the joint energy in a global optimization MRF framework, both of these two methods rely on corresponding unique saliency-based strategies as initialization to help address robust object extraction problem.

For better understanding, we design a label merging strategy to portray the link properties between the local pixel and global region information in the Cross-Region potential. The advantage of using this link term is that we do not need to train many classifiers which depend largely on the combinational features. Further, the concept of hierarchy segmentation [21, 22, 28, 30, 39] has aroused researchers' attention in recent years, and most existing hierarchy segmentation methods use the scale-space structure of an image to explore the similarity of pixels at multiple scales. Different from solving the segmentation problem at multiple scales, we utilize serial semantic labels to preserve the hierarchical structure of segmentation results. Based on this, our approach unexpectedly preserves the object shapes in natural images as in [2]. Different from Lin and Wang's approach in [2] to recognize the object shapes by utilizing the node and layout of the And-Or graph model, the hierarchical scale information in our method is aggregated by a hierarchical label alignment mechanism in the Cross-Scale potential to exploit the topological properties of labels on different scales.

The key contributions of this paper are as follows: (1) We present a novel approach to capture the context, layout, and scale information in a given image and achieve effective segmentation using the LSI framework; (2) we formulate the segmentation problem as a label inference problem in the label space to reason for the label of each pixel; (3) Cross-Region and Cross-Scale potentials

are designed for the label inference algorithm to derive the final segmentation. Hence, the proposed approach can easily be expanded for Object-Class Segmentation and Semantic segmentation [7, 8] tasks based on the properties of the objects.

The remainder of the paper is organized as follows: In Section 2, existing segmentation methods are presented. In Section 3, the proposed LSI framework is elaborated. Then, experiments are described in detail in Section 4. Finally, Section 5 concludes the paper with directions for future work.

2. Brief Review of Related Work

There are clear distinctions between the proposed LSI approach and previous works. In this section, we briefly review the existing works and comment on the advantages of our method.

(1)Problem Formulation and Transformation

The majority of recent works in this area focus on solving a graph-based problem [16, 17], clustering-based problem [18, 19], or a hierarchy-based problem [21, 22] and regard the general image segmentation problem as Object-Class segmentation and Semantic segmentation. Our method treats the task as a label inference problem which is different from the aforementioned ideas. In the graph-based segmentation methods, the problem is represented as graph partition, and the global information of an image is usually utilized. However, the extraction and employment of global information is very limited and results in high storage requirement inevitably. In comparison, our model involves the rich information from local and global context, as well as layout and scale information. To reduce the storage requirements, we only catch the image local information in sub-image level, and infer the final segmentation in image level. The clustering-based segmentation approach [18] aims to group pixels with similar patterns into the same cluster by maximizing the inter-cluster dissimilarity and the minimizing the intra-class similarity. This method, however, has a intrinsic limitation that the number of clusters are unknown. In our method, although

the number of segment classes is still unclear, we can estimate it approximately. To obtain the label of a pixel, we make use of the clustering algorithm proposed in [29] to capture the local information in label formation phase. The state-of-art hierarchical segmentation method [22] is commonly based for the analysis of local feature cues, which are restricted and are focused on the segmentation results in terms of a hierarchy, especially in Ultra Contour Map (UCM) algorithm [21].

(2)Model Categorization and Comparison

For natural images, there exists a vast amount of sophisticated models in the aforementioned three segmentation methods. The work in [12] proposes an inhomogeneity-embedded active contour (InH-ACM) for natural image segmentation by minimizing the energy of global color coherence and local inhomogeneity consistency. However, this method needs a saliency-inspired framework to start the evolution of locating the initial contour for InH-ACM, while our method avoids the subsequent steps on the contour which may be imprecise. The work in [34] proposes an adequate variational segmentation model for segmentation of images perturbed by arbitrary noise models, which is intend to be applied to real application data from biomedical images. Our method is designed for natural image segmentation without losing the generality of being applied in other domains.

The proposed LSI framework utilizes the MRF-inspired model to infer the semantic label. Feng and Jia [35] address the problem of self-validated labeling of Markov random fields, by treating the whole image as a single segment with three concrete graph cuts algorithms, and converse when the energy stops decreasing. In the graph cuts method, each pixel serves as a node in graph theory, whereas our method treats each subregion in the label formation phase as a node in the label space. One of the advantages about our label representation is the fact that the storage consumption is greatly reduced. Sfikas and Nikou [33] propose a new Bayesian model for image segmentation by defining local and global weights with a spatial variant and an MRF edge-preserving smoothing prior, respectively. Although this method is used for natural image segmentation, it

still begins with an existing superpixel initialization. In comparison, we design a CLDC algorithm to extracts the sub-tags for label inference. Therefore, in sub-regions, the proposed method catches more abundant local cues than most significant superpixel initialization. This is because we extract the tags from sub-images of smaller size in the label formation phase. With the constraints of a MRF, Lin and Liu [32] present a framework that incorporates shape and structure information into a sketch graph, which applies to object categorization instead of image segmentation, while our method can preserve the shape information by utilizing the scale information, and the experimental results in Figure 7 can verify the effectiveness.

(3)Hierarchical Structure Description

A typical segmentation algorithm with the idea of hierarchy, designed by Arbelaez and Maire [21], consists of generic machinery for transforming the output of any contour detector into a hierarchical region tree with low time efficiency. Based on that, Donoser and Schmalstieg [22] propose a hierarchical image segmentation model. It is described as a coarse-to-fine structure to exploit different levels of contextual information preferring to predict local gradients for each pixel in a test image. Zheng and Cheng [30] develop a hierarchical model to incorporate region-level objects and attribute information in the semantic segmentation, such as Wood, Cotton, etc. The serial labels used in our approach are not semantic concepts, since most approaches based on MRF are designed to focus on reasoning the relationship of existing labels. Additionally, Arbelaez and Malik [31] propose a unified approach for bottom-up hierarchical image segmentation, which is later used for recognition. In the method, combinational space is explored for integrating the multi-scale regions into highly-accurate object candidates instead of general image segmentation. In contrast to these approaches, we take advantage of the scale information of labels in the label scale space rather than in the image scale space.

3. Latent Semantic Inference of Serial Label

In this section, we introduce the segmentation framework for the Latent Semantic Inference of Serial Label (LSI). For a given image, we aim to segment it into perceptual regions by assigning each pixel an unspecified label. Here, the unspecific label is defined as a serial number label instead of the kind of semantic objects. Now we describe the solution of our LSI framework with a visual effect in Figure 1 using Image-143090 in BSD300 as an example. Especially, we take full advantages of the local pixel information, global region information, and scale information.

In the latent semantic inference framework, the input image is S1. First, S1 need to be down sampled twice. With the proposed color and location density clustering algorithm named CLDC, we get the label of each scaled image in the label formation phase. Furthermore, label merging strategy and label alignment mechanism are integrated into the MRF model in the following label inference phase. Finally, we get the segmentation results with the proposed LSI.

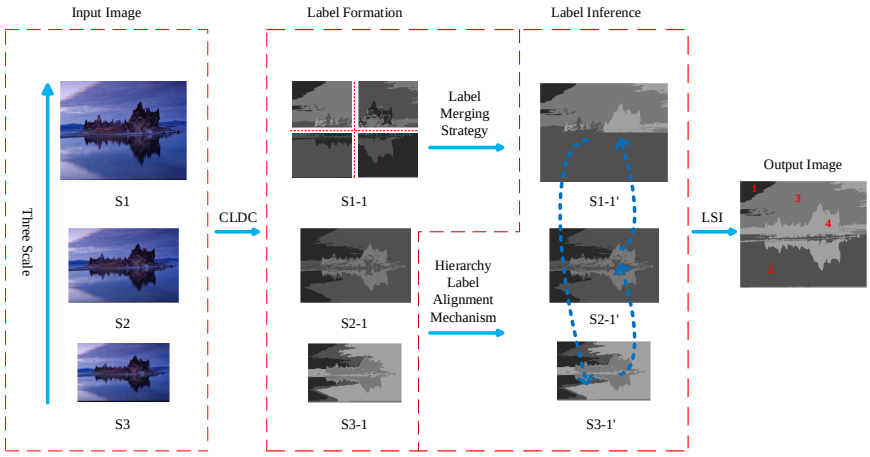

Figure 1: Latent Semantic Inference Framework for Segmentation.

3.1. Color and Location Density Clustering

The main idea of label formation phase is to obtain the labels in the inference process by computing the similarity of all pixel pairs in the whole image. To this

end, a color and location density clustering algorithm, named CLDC, is briefly introduced. In our algorithm, instead of starting with an over-segmentation [23, 24, 25] for the whole image, i.e. a set of superpixels [20] partially adhering to local boundaries, we split the given image into quarters along the red lines to extract more detailed local pixel information from each independent part, which illustrated in Figure 2. The quadrant structure supports well the label merging strategy.

We take each independent part as a sub-image. From the psychological point of view [26, 27], after splitting an image into quarters, the bottom-up and left-right searching method is in accord with the human beings cognitive habits. Furthermore, to avoid the potential memory consumption issue, we perform CLDC algorithm in each sub-image instead of the whole image.This brings up two problems: (1) How to measure the similarity of pixels within the sub-image? (2) How to assure the integrity of superpixels across different sub-image?

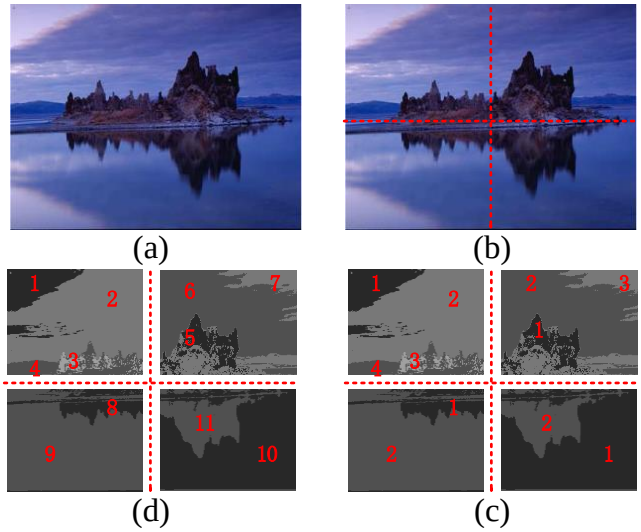

Figure 2: Labels formation:(a) original image, (b) sub-images, (c) serial labels in sub-image, (b) sequence labels in image.

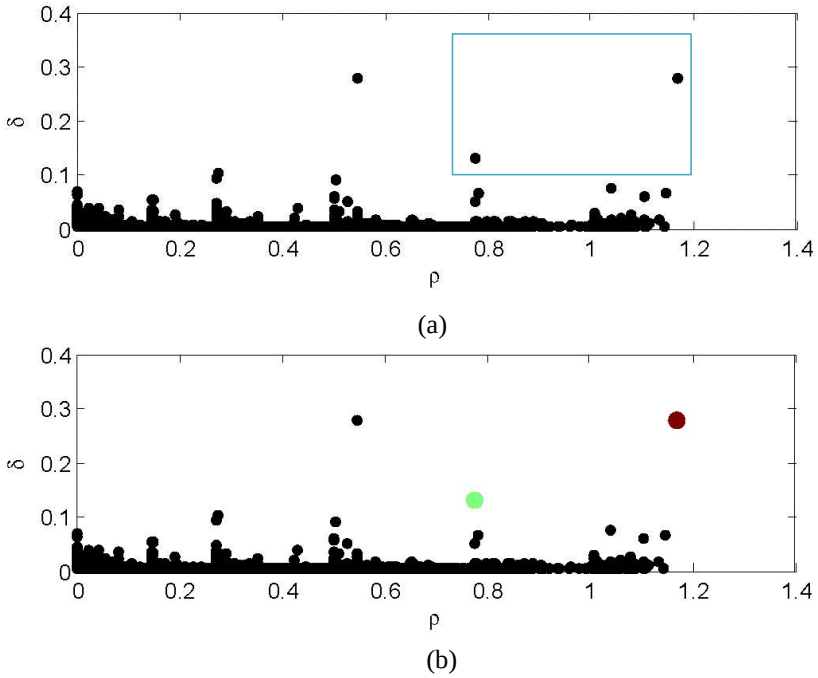

Figure 3: The selection of ρ and δ in the label formation phase. We select ρ and δ manually. (a) is the distribution of ρ and δ decision graph. If the value of selected ρ is large than half of the maximum value, we draw the rectangle from the top right corner in (a). (b) shows the colored cluster centers.

To solve the first problem, for each sub-image, each pixel is treated as a node according to the Graph-Based Segmentation algorithm. We compute the similarities of all pixel pairs in the sub-image. Different from Graph-Based algorithm, we select method in [29], which leads to good effect in cluster analysis, to do the sub-image clustering. For the reason that [29] can not be directly applied to traditional image segmentation task due to its high storage consumption and the irregularity of pixel distribution, we use sub-images to reduce the storage consumption and find a mode of normalizing the pixel similarity to 2D data distribution.

In the label formation phase, to get the label presented in Figure 2(c), first and foremost, each pixel of the sub-image is represented as a five-dimensional

vector $p_i = [l_i, a_i, b_i, x_i, y_i]$ in the CIELAB color space, where $[x_i, y_i]$ are the space coordinates in the sub-image. Then we use Euclidean color and spatial distances to measure the similarity of all pixel pairs $D_{p_{ij}}$ defined as follows:

$$\begin{aligned} dc_{ij} &= \sqrt{(l_i - l_j)^2 + (a_i - a_j)^2 + (b_i - b_j)^2} \\ ds_{ij} &= \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \\ D_{p_{ij}} &= dc_{ij} + ds_{ij}, \end{aligned} \tag{1}$$

where dc_{ij} is the lab color distance and the ds_{ij} is the spatial distance. We expect that the superpixels(i.e. cluster centers) are surrounded by pixels with lower local density and the density between clusters is relatively higher. Therefore, using the clustering method similar to [29], for each pixel, we compute two values: local density ρ_{p_i} and distance δ_{p_i}.

The local density ρ_{p_i} is given as:

$$\rho_{p_i} = \sum_{p_j} \chi(D_{p_{ij}} - t), \tag{2}$$

where $\chi(x) = 1$ if $x < 0$, or $\chi(x) = 0$, otherwise, and t is a threshold of similarity distance. According to Equation 1, we set t the mean value of $D_{p_{ij}}$. In this way, we can transform the similarity matrix of pixel in sub-image into standards 2D data points distribution to find the position of the cluster center. The distance δ_{p_i} denotes the minimum similarity distance between p_i and p_j:

$$\delta_{p_i} = min_{\mathbf{j}}(D_{p_{ij}}), \tag{3}$$

with the condition $\rho_{p_j} > \rho_{p_i}$. With the ρ and δ, we can find the cluster centers as those with large values of both the two. The selection of ρ and δ is presented in Figure 3.

Finally, we sort the serial label in all four sub-images. From Figure 2(d), we can observe that there is no link between the sorted labels. However, along the red line we can find some label pairs like $(2, 6)$, $(8, 11)$, $(9, 10)$ denote different classes while in fact they should belong to the same region. Therefore, we will solve this in the design of our Cross-Region potential of the $Cross - R\&S$ model in label inference phase.

3.2. Cross − R&S Model

One key challenge in the LSI framework is how to utilize a rich mixture of basic image information such as color, texture, and middle-level features. To address this challenge, we develop the $Cross - R\&S$ model which is able to appropriately integrate multiple information into a MRF-inspired model. In this model, we treat the segmentation task as a label inference problem with a MRF-inspired model but rather different from traditional MRF settings.

Here the key differences from the traditional MRF method [10, 31, 35] include (1) we operate label inference in the label space instead of pixel space, (2) the labels are obtained by our CLDC algorithm rather than traditional superpixel-based methods [21, 23, 24, 25, 31], and (3) the label inference associates a label merging strategy in the unary term with a hierarchy label alignment mechanism in the pairwise term.

Similar to that of [10], the energy function of region labels in $Cross - R\&S$ Model is defined as:

$$E(l) = \sum_{i=1}^{M} \psi_i(l_i) + \lambda \sum_{i,j=1}^{M,N} \phi_{ij}(l_i, l_j), \tag{4}$$

where M is the total number of labels, and N is the total number of neighborhood labels. In this energy function, the Cross-Region potential is defined as ψ_i and the Cross-Scale potential is defined as ϕ_{ij}. In the Cross-Region potential, based on labels obtained in the label formation phase, we introduce a label merging strategy to catch the global region information of an image. In the Cross-Scale potential, a hierarchy label alignment mechanism is presented to grasp the layout information of image. On the basis, we propose an overall inference algorithm to achieve our segmentation. Finally, our goal is to score the entire description of the image labels by iteratively minimizing the energy function.

3.2.1. Cross-Region Potential

Once subimages have been segmented into independent subregions which are indicated as R_i^j, where $i = \{1, 2, 3, 4\}$, and j is the order of cluster centers, i.e. the serial number of labels in the i-th sub-image, we aim to design a strategy to

merge these subregions into continuous meaningful regions R_k across so as to capture the global region information of the sub-images. Now we describe how to assure the integrity of superpixels across different sub-images mentioned in the label formation stage.

For each obtained independent subregions, we define a label merging strategy

$$f^l : R_i^j \rightarrow R_k \tag{5}$$

where l is the index for discrete label presented in Figure 2(d), and k is the region label after the subregions are merged. The aim is to eliminate the incoherence of these subregions across the red lines. The merging strategy is summarized in Algorithm 1.

Algorithm 1 Label merging strategy

Input: Subregion R_i^j
Output: Region R_k
Count the sum N number of the labels in all sub-regions R_i^j
Compute the number of pixels N_{pair} and patch similarity S_{pair} between two adjacent sub-regions along the red horizon and vertical lines in the middle of the image
/ Subregions Merge: */*
repeat
  for $num = N$ **do**
    if $(N_{pair} > 5) \bigcap (S_{pair} < \sum_i max(\sum_j \delta(R_i^j)))$ **then**
      $R_{num} \rightarrow R_k$
    end if
  end for
until $num = 1$

In the function f^l, we set $l = l_1, l_2, ...l_n$, where n is the sum of cluster numbers in all four subimages, then the strategy degenerates from $R_i^j \rightarrow R_n$, so next step is to merge R_n into R_k. Based on this, we consider two quantifications as the

norm of sub-region integration: the number of pixels between two adjacent sub-regions N_{pair} and patch similarity between two adjacent S_{pair} along the red lines in the middle of the image.

First we count the number of pair number of these pixels along the red lines in the middle of the image. If N_{pair} is smaller than 5, we empirically regard the two subregions have no correlation. Then, we use the first formula of Equation 1 to compute similarity of the counted pixel pairs, thus confirming whether the two adjacent subregions are with high correlation. The value of pixel pair similarity defined as S_{pair}, is subject to the following conditions:

$$S_{pair} < \sum_{i} max(\sum_{j} \delta(R_i^j)) \tag{6}$$

Finally, according to the above constraints, all subregions R_n are fused into R_k until the clusters in subimages are run through.

According to our strategy, we define our Cross-Region potential ψ as:

$$\psi_i(l_i) = -log(\sum_{i} max(\sum_{j} \delta(R_i^j)) - (\sum_{i}\sum_{j} S_{pair})/k), \tag{7}$$

where S_{pair} is subjects to the condition given in step 8 in Algorithm 1.

3.2.2. Cross-Scale Potential

We construct an image pyramid with 3 scales by subsampling the original given image, producing three images for each original, with 481*321, 241*161, 121*81 pixels. These three scale images are referred to as S1, S2, S3 respectively. As mentioned in Section 1, we perform operations of label inference in the label-scale space rather than image-scale space. The fundamental step in Cross-Scale potential is to get the labels in different scales. Therefore, each pixel of scaled image S2 or S3 is represented as a five-dimensional vector $p_i = [l_i, a_i, b_i, x_i, y_i]$ in the CIELAB color space, then we use Euclidean color and spatial distances to measure the similarity of all pixel pairs $D_{p_{ij}}$, and define the local density ρ_{p_i} in the same way as in the label formation phase. These steps produce a set of Cross-Scale potential: S1-1, S2-1,S3-1.

After three fixed scale images are segmented, we design a hierarchy label alignment mechanism to describe the labels on three scales, which is elaborated in Figure 4.

Our main idea is to confirm the adjacent labels in comparison to different scale input images. In this mechanism, we first empirically consider that S3-1 entails the most layout information and S2-1 has more layout information than S1-1. For this reason, if the label of S3-1 occurs as that of S2-1 and S1-1, we set S3-1's label as the new label for S2-1. Meanwhile, the value of S_{pair} modifies with the new label in the S1-1 by adopting the label merging strategy in the Cross-Region potential. Though this mechanism, we can get new S1-1' and S2-1'.

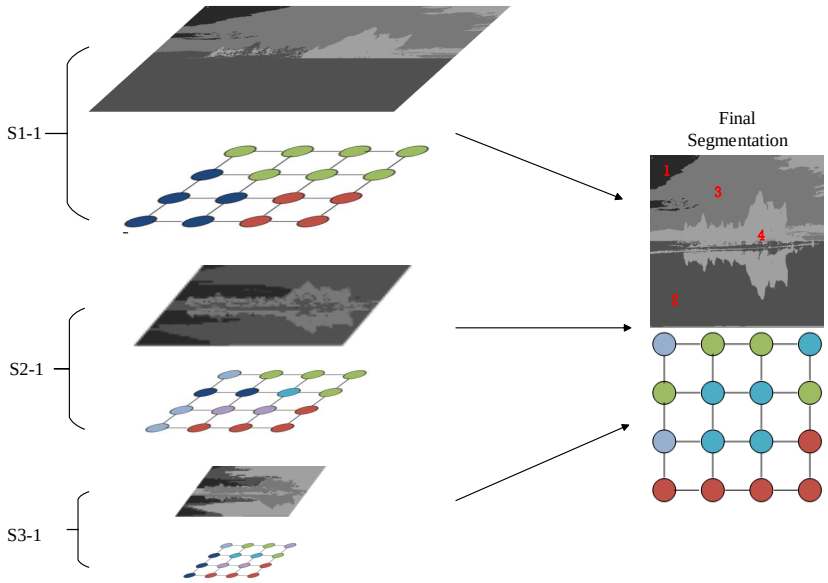

Figure 4: Label inference in Cross-Scale Potential. S1-1, S2-1, S3-1 are scaled images. If two images in three are with the same label, the rest image updates the label in the same position, then we compute the cross-scale potential. If the label in S1-1 changes, the cross-region potential is modified by reusing the label merging strategy.

Therefore, we define the Cross-Scale Potential ϕ_{ij} as:

$$\phi_{ij}(l_i, l_j) = -log(Num_{S3-1}(l_i, l_j) - Num_{S2-1}(l_i, l_j)), \tag{8}$$

where $Num_{S3-1} - Num_{S2-1}$ is the sum of different labels between S2-1 and S3-1.

3.2.3. Inference

The overall inference algorithm of $Cross-R\&S$ model is described as: First, we obtain the labels of an image at three different scales: $S1- > S1-1$, $S2- > S2-1$, $S3- > S3-1$, and initialize the unary potential ψ using our Label Merging Strategy with S1-1 and the pairwise potential ϕ with S2-1 and S3-1, with which the total energy E can be computed; Next, given potential ψ, we compare the second scaled segmented image S1-1 and S3-1, then update the Cross-Region potentials ψ; and compare S2-1 and S3-1, then update the pairwise potential ϕ; Finally, we compute the new total energy E_{new}, if $E_{new} < E$, we set $E = E_{new}$. The algorithm iterates till convergence is reached.

4. Experiment

In this section, we show the experimental results of the proposed **LSI** on Berkeley Segmentation Dataset. The proposed **LSI** includes two phase. The label formation takes from 40 to 60 seconds, and the label inference phase costs no more than 0.5 second for one image. We also evaluate the performance and compare the accuracy with recent popular approaches.

Berkeley Segmentation Dataset: The Berkeley Computer Vision Group collected 12,000 hand-labeled segmentations of 1,000 Corel dataset images from 30 human subjects. Half of the segmentations were obtained from presenting the subject with a color image; the other half from presenting a grayscale image. The public benchmark based on this data consists of all of the grayscale and color segmentations for 300 images. The images are divided into a training set of 200 images, and a test set of 100 images.

To demonstrate the effectiveness of our approach, we compare its performance with the well-known Ncuts [16], Meanshift [18], and SLIC [20] segmentation methods which are reviewed in Section 2. Figure 5 provides results of LSI and three aforementioned methods on several images from the BSD. In Figure

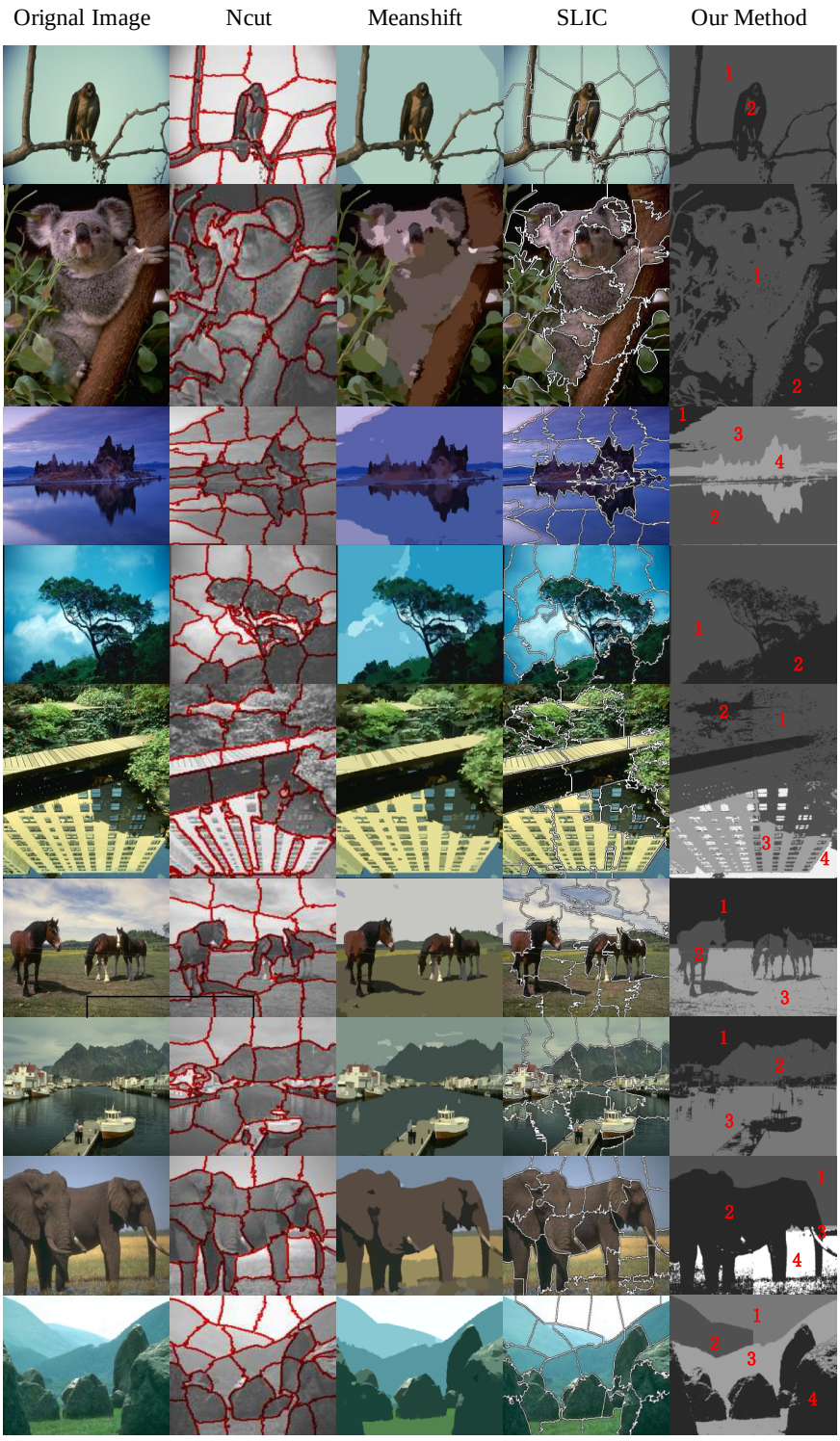

Figure 5: Given an image, we segment it into perceptually semantic regions.

5, we set the number of segments in Ncut, Meanshift, SLIC to be uniformly 30 which is empirically decided as a good number in average. We can observe that in the image of the third row, sky, cloud, mountain and water are segmented neatly, while in Ncut method, regions are not meaningful. If the ultimate task is to do recognition, various kinds of combination models or discriminative classifiers can be used to further integrate these segments. From the sixth and eighth row in Figure 5, we can observe that the horses and elephants are segmented from the complex background. Therefore, although a few objects are wrongly categorized into one class, for example, the shadow of a horse is segmented into the horse, the segmentation by LSI method is useful for object-class segmentation, semantic segmentation and scene understanding tasks.

Table 1: Region Benchmark on BSD300

BSD300		
	Segmentation Covering	Rand Index
Groundtruth	0.73	0.87
Ncut	0.53	0.79
Meanshift	0.58	0.80
SLIC	0.63	0.83
LSI	**0.59**	**0.80**

Instead of using the boundary methodology which supports contour detectors over segmentation, we evaluate our method and Ncut, Meanshift, SLIC approach with region-based criteria which includes Rand Index [21] and Segmentation Covering [36, 37, 21]. In Rand Index criteria, the Probabilistic Rand Index is defined as:

$$PRI(S, G_k) = 1/T \sum_{i<j} [c_{ij} p_{ij} + (1 - c_{ij}(1 - p_{ij}))],$$

where S and G are the test and ground-truth segmentation, and c_{ij} is the event that pixels i and j have the same label pairs, p_{ij} is its probability, and T is the total number of pixel pairs. In Segmentation Covering criteria, the covering

between two different segments S and S' is defined as:

$$O(S' -> S) = 1/N \sum_{R \in S} |R| \cdot max_{R' \in S'} |R \cap R'|/|R \cup R'|,$$

where N is the total number of pixels in the image. Table 1 presents region benchmarks on the BSD.

From Table 1, we can observe that the proposed LSI achieves satisfactory results, just slightly inferior to SLIC. Although region-based criteria provides an important measure for segmentation, the advantages of LSI is not fully demonstrated through these measures. The key idea of LSI is to understand the segmentation results in a way that is similar to human visual observation. The main goal of this approach is to achieve visual segmentation results that can be meaningfully interpreted as they are analysed by human beings, instead of producing relatively independent regions. Therefore, we analyse segmentation results from two additional aspects in terms of numbers of segments and meaning of the segmented regions.

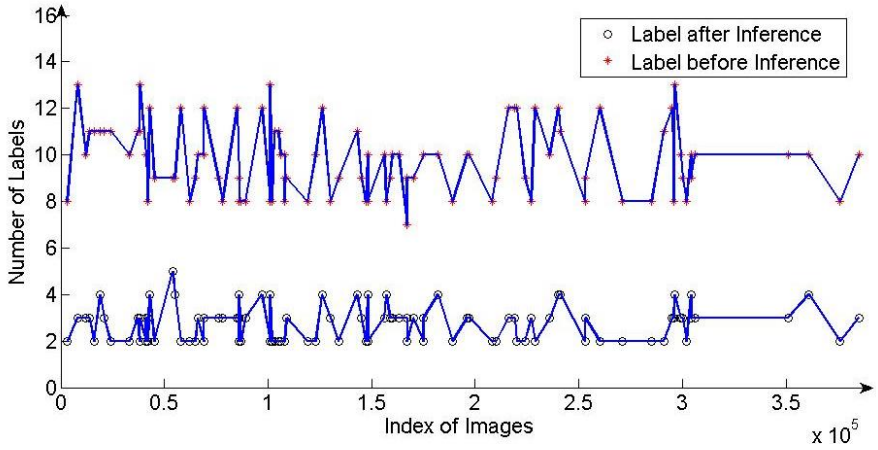

Figure 6: Label statistics before and after inference.

To test the efficiency of the estimated number of segments, we first give a quantitative statistics of feasible existing object class as in the given image. The results presented in Figure 6 show that for natural images, in most cases, they

are segmented into 2-4 meaningful regions with our LSI approach. For a human being, it is fairly enough to recognize the shape of segment regions with 2-4 meaningful regions. We can see from Figure 6 that the number of regions in most natural images is reduced from 8-14 to 2-4 by the label inference. This matches with the fact that human interpret images into 2 meaningful regions. It can be observed in the Ncut segmentation image regions are not coherent with human visual perception because they are not semantically meaningful. We can see it clearly from the images in the last column in Figure 7 with concise background are intend to be segmented into two classed, i.e. foreground and background, and images with complex background generate more regions than ones with concise background. The background here is an amorphous object [38], because it depends on how ambiguous and complex the background is intend to be.

To understand the meaning of segmented regions, we observe that from the second image in the first two rows in Figure 5, these two results can be intuitively understood as 'an eagle standing on a branch' and 'a koala holding a branch'. Moreover, in Ncut, Meanshift, and SLIC methods, the number of segments needs to be specified manually, which means the performance heavily relies on the specified number of segments, and unsuitable number of segments may lead to segmentation that are not semantically meaningful to human.

Further, we compare our segmentation results with the three other methods using different numbers of regions in Figure 7. In Figure 7(a), the number of segments in Meanshift is set to 5, and we can see that Meanshift intends to produce smaller regions which are confusing. In Figure 7(b), we use two setups for the number of regions for Ncut. In the first line in Figure 7(b), when the number of regions is set to be the same with those LSI obtains, Ncut produces similar results to our method. In the second line when the approximate number of segments in Ncut method are used, the quality of our segments is superior to Ncut. In Figure 7(c), the number of segments in SLIC is set to 20 and 200. When the number of segments is 20, we can see that our method can preserve more local information in the superpixels. In addition, although the local information

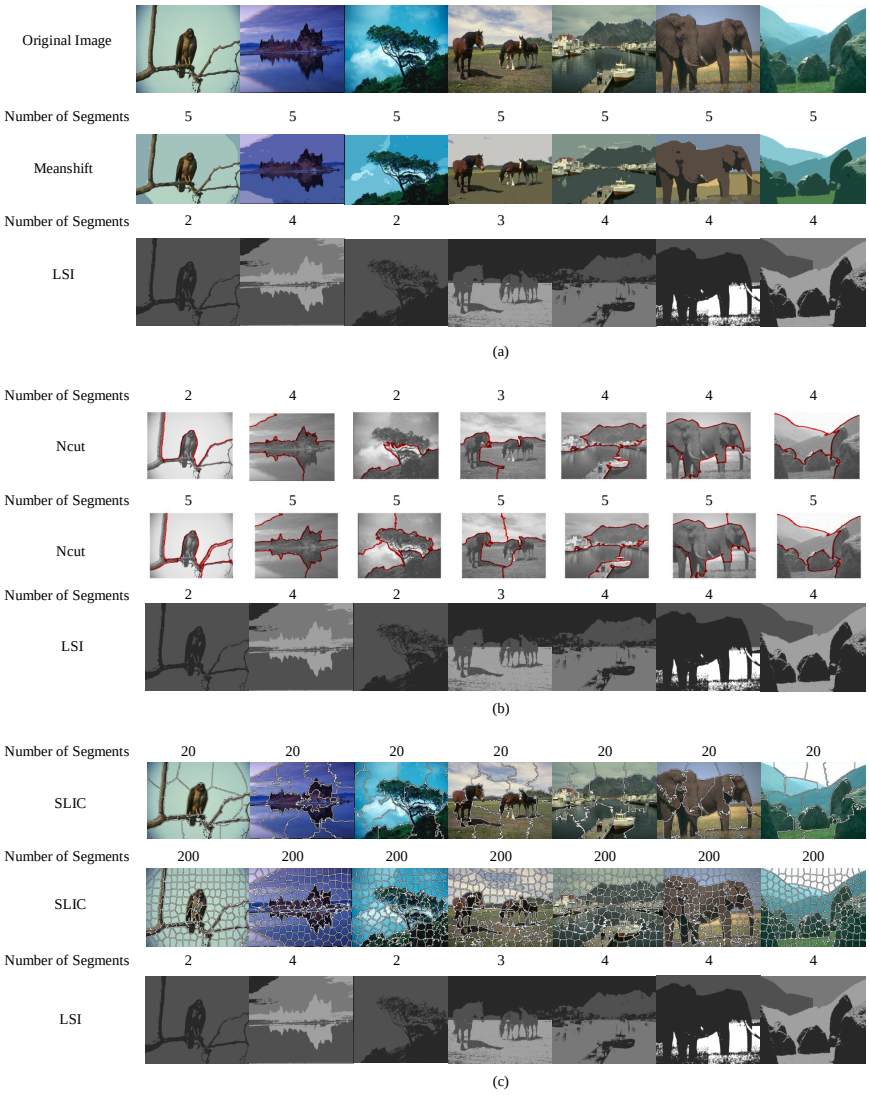

Figure 7: Comparison with different number of segments.

in SLIC can be more abundant when the number of segments is 200, our method can catch much more global information and master the meaning of regions. Therefore, analysing the number of segments and the meaningfulness of the segmented regions, we can observe that our method outperforms the three other methods, and can preserve the object shape at large extent.

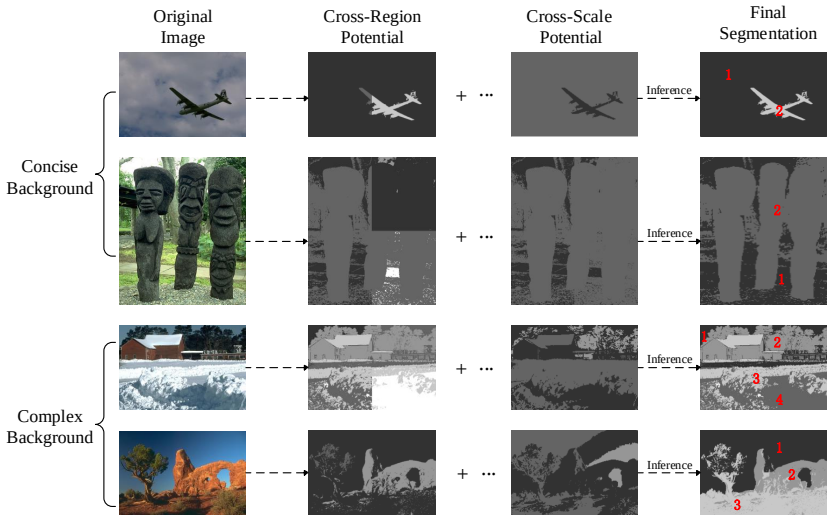

Figure 8: Given an image, we catch the local and global layout information of the original image through the inference stage.

More specifically, we analyse the segmentation performance of two potentials in Figure 8. The second and third columns from the left show preprocessed images: initial images produced by label merging strategy and initial images produced by hierarchy label alignment mechanism. We can see clearly that images in second column with concise background are almost merged together with less of layout information, while for images with complex background the layout information is captured at the cost of a large number of missing of local information. Therefore, we bridge these two potentials together to infer regions. It can be observed in column 4 that each segment accounts for a meaningful object. For example, for image in the third row, label 1 represents sky, label 2 shows trees, label 3 shows snow, and label 4 is the shadow of snow. The

Ncut, Meanshift, and SLIC methods are unable to preserve the integrity of semantic object in the segmentation phase. Therefore, such methods usually can only serve as an initialization step in semantic segmentation or object-class segmentation approaches, while our method provides a viable alternative for these application.

Observing that each segment is always a meaningful object in the the rightmost column, we think of the serial label 1 of image in the third row as sky, label 2 as trees, label 3 as snow, and label 4 as shadow of snow in the psychological perception, while the Ncut, meanshift, and SLIC methods are unable to preserve these meaningful information in the segmentation phase. Therefore, most of semantic segmentation or object-class segmentation approaches select the aforementioned three methods as an initialization rather than manipulating features in pixel-level directly, while our method provides a viable alternative for these applications.

5. Conclusion

The goal of this paper is to explore segmentation approaches for partitioning the given image into perceptually meaningful regions. We design a LSI framework using a $Cross - R\&S$ model including a Cross-Region potential and a Cross-Scale potential to integrate detailed image information for the final segmentation. It is worth noting that we regard image segmentation as label inference problem. The label is obtained by a label formation phase instead of the recent popular superpixel initialization approach, and a label merging strategy is then performed to grasp the local and global information of regions. Then, utilizing the layout information at different scales, a hierarchy label alignment mechanism is formed to infer the final segmentation. In addition, according to the final result, our LSI is superior in terms of object-class segmentation, semantic segmentation and scene understanding. These advantages will be explored in semantic recognition and understanding tasks in our future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61370149, in part by the Fundamental Research Funds for the Central Universities(ZYGX2013J083), and in part by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] S. Duffner, J. Odobez, "Leveraging colour segmentation for upper-body detection," in *Pattern Recognition*, vol. 47, no. 6, pp. 2222-2230, 2014.

[2] L. Lin, X. Wang, W. Yang, and J. Lai, "Discriminatively trained And-Or graph models for object shape detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 959-972, 2015.

[3] S. Roy, S. Mukhopadhyay, M. K. Mishra, "Enhancement of morphological snake based segmentation by imparting image attachment through scale-space continuity," in *Pattern Recognition*, vol. 48, no. 7, pp. 2222-2230, 2015.

[4] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *Proceedings of IEEE International Conference on Computer Vision*, 2011.

[5] L. Lin, P. Luo, X. Chen, and K. Zeng, "Representing and recognizing objects with massive local image patches," in *Pattern Recognition*, vol. 45, no. 1, pp. 231-240, 2012.

[6] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[7] F. Li, J. Carreira, G. Lebanon, and C. Sminchisescu, "Composite statistical inference for semantic segmentation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[8] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. D. ourdev, and J. Malik, "Semantic segmentation using regions and parts," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[9] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," in *IEEE Transactions on Image Processing*, in press, 2015.

[10] H. Myeong, J. Y. Chang, K. M. Lee, "Learning object relationships via a graph-based context model," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[11] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2-23, 2009.

[12] L. Dai, J. Ding, J. Yang, "Inhomogeneity-embedded active contour for natural image segmentation," in*Pattern Recognition*, vol. 48, no. 8, pp. 2513-2539, 2015.

[13] C. Galleguillosy, B. McFeey, S. Belongiey, and G. Lanckriet, "From region similarity to category discovery," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

[14] Y. Zhang, M. Abdel-Mottaleb, Z. He, "Unsupervised segmentation of highly dynamic scenes through global optimization of multiscale cues," in*Pattern Recognition*, vol. 48, no. 11, pp. 3477-3487, 2015.

[15] L. Zhu, Y. Chen, Y. Lin, C. Lin, and A. Yuille, "Recursive segmentation and recognition templates for image parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no.2, pp. 359-371, 2012.

[16] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.

[17] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167-181, 2004.

[18] D. Comaniciu and P. Meer, "Mean Shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, 2002.

[19] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proceedings of European Conference on Computer Vision*, 2008.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274-2282, 2012.

[21] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898-916, 2011.

[22] M. Donoser and D. Schmalstieg, "Discrete-continuous gradient orientation estimation for faster image segmentation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[23] S. Kim, S. Nowozin, P. Kohli,C. D. Yooin, "Higher-order correlation clustering for image segmentation," in *Advances in Neural Information Processing Systems*, 2011.

[24] D. Hoiem, A. A. Efros, M. Hebert, "Recovering occlusion boundaries from an image," *International Journal of Computer Vision*, vol. 91, no. 3, pp. 328-346, 2011.

[25] X. Liu, L. Lin, and A. L. Yuille, "Robust region grouping via internal patch statistics," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[26] M. B. Mathur, D. B. Reichling, "Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley," *Cognition*, vol. 146, pp. 22-32, 2016.

[27] L. Oksama, J. Hyona, "Position tracking and identity tracking are separate systems: Evidence from eye movements," *Cognition*, vol. 146, pp. 393-309, 2016.

[28] Z. Ren and G. Shakhnarovich, "Image segmentation by cascaded region agglomeration," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[29] A. Rodriguez, AlessandroLaio, "Clustering by fast search and find of density peaks," *Science*, Vol. 344, no. 6191, pp. 1492-1496,2014.

[30] S. Zheng, M. Cheng, J. Warrell, P. Sturgess, V. Vineet, C. Rother, P. H. S. Torr, "Dense semantic image segmentation with objects and attributes," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[31] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, J. Malik, "Multiscale combinatorial grouping," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[32] L. Lin, X. Liu, S. Peng, H. Chao, Y. Wang, and B. Jiang, "Object categorization with sketch representation and generalized Samples," in *Pattern Recognition*, vol. 45, no. 10, pp. 3648-3660, 2012.

[33] G. Sfikas, C. Nikou, N. Galatsanos, and C. Heinrich, "Majorization-Minimization mixture model determination in image segmentation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

[34] D. Tenbrinck, X. Jiang, "Image segmentation with arbitrary noise models by solving minimal surface problems," in *Pattern Recognition*, vol. 48, no. 11, pp. 3293-3309, 2015.

[35] W. Feng, J. Jia, and Z. Q. Liu, "Self-validated labeling of markov random fields for image segmentationin," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1871-1887, 2010.

[36] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman "PASCAL 2008 Results," http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html, 2008.

[37] T. Malisiewicz and A.A. Efros, "Improving spatial support for objects via multiple segmentations," in *Proceedings of British Machine Vision Conference*, 2007.

[38] S. Zheng, M. Cheng, J. Warrell, P. Sturgess, V. Vineet, C Rother, P. Torr, "Dense Semantic Image Segmentation with Objects and Attributes," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[39] L. Dong, J. Su, E. Izquierdo, "Scene-Oriented Hierarchical Classification of Blurry and Noisy Images," in *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2534-2545, 2012.