A Probabilistic Framework for Unsupervised Evaluation and Ranking of Image Segmentations

Mustafa Jaber(1), Sreenath Rao Vantaram(1), and Eli Saber(1,2)

1Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623 USA
2Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623 USA
{mij7272, sxv9436, esseee}@rit.edu

Abstract- In this paper, a Bayesian Network (BN) framework for unsupervised evaluation of image segmentation quality is proposed. This image understanding algorithm utilizes a set of given Segmentation Maps (SMs) ranging from under-segmented to over-segmented results for a target image, to identify the semantically meaningful ones and rank the SMs according to their applicability in image processing and computer vision systems. Images acquired from the Berkeley segmentation dataset along with their corresponding SMs are used to train and test the proposed algorithm. Low-level local and global image features are employed to define an optimal BN structure and to estimate the inference between its nodes. Furthermore, given several SMs of a test image, the optimal BN is utilized to estimate the probability that a given map is the most favorable segmentation for that image. The algorithm is evaluated on a separate set of images (none of which are included in the training set) wherein the ranked SMs (according to their probabilities of being acceptable segmentation as estimated by the proposed algorithm) are compared to the ground-truth maps generated by human observers. The Normalized Probabilistic Rand (NPR) index is used as an objective metric to quantify our algorithm’s performance. The proposed algorithm is designed to serve as a pre-processing module in various bottom-up image processing frameworks such as content-based image retrieval and region-of-interest detection.

Keywords: Segmentation Evaluation, Bayesian Networks, Image Understanding.

I. INTRODUCTION

Many image segmentation algorithms have been proposed in the recent years to group image pixels yielding semantically meaningful maps. Different mathematical and physical theories have been employed to develop these techniques as summarized in the surveys [1] and [2]. The majority of the state-of-the-art image segmentation techniques yield in a single map for any target image with the assumption that it is the optimal segmentation under certain conditions. Spatial and spectral features are used as optimization criteria to generate the best (optimal) segmentation map. This may include empirically setting some thresholds or setting a number for expected segments in the output map.

Furthermore, the segmentation algorithms in the literature are usually evaluated and benchmarked against other algorithms on a limited set of test images. This comparison may yield that a certain methodology of segmentation performs better on a set of images but this does not mean that this “winner” algorithm will definitely generate the optimum Segmentation Map (SM) for any target image. In this paper, we try to overcome these limitations by developing a framework to identify a number of semantically acceptable segmentation maps of a test image, and estimate the probability of each map of being the optimal segmentation of the target image. The probability values are used to rank the segmentations in accordance to their usability.

Methodologies for evaluating the performance of image segmentation algorithms are having more attention in the recent years. They are essential to systematically determine the merit, worth, and significance of the segmentation techniques. Objective and subjective segmentation evaluation methods are found in the literature and are summarized in the survey by Zhang et al. [3]. More resent segmentation evaluation techniques are found in [4], [5]. The method in [4] is based on genetic algorithms which were used to minimize feature criteria over different segmentation maps. The search space is defined as the label of each pixel of the test image. Furthermore, the technique proposed in [5] utilized energy formulation to generate multi-scale criteria. However, the level of detail expected by the evaluation criterion is left to the user to select.

In this paper, an image understanding algorithm for ranking segmentation maps of an arbitrary image into different levels according to their usefulness is proposed. An “acceptable” segmentation is a map that depicts a semantically meaningful partitioning of image regions. The proposed algorithm has two phases: training phase which utilizes ground truth segmentation maps to build a Bayesian Network (BN) and a testing phase. The proposed algorithm includes modules for image segmentation, feature extraction, discretization, and probabilistic reasoning. The proposed algorithm generates multiple segmentation maps utilizing the algorithm in [6] and ranks them according to their estimated probability of being acceptable in image processing and computer vision systems.

The remainder of this paper is organized as follows. The process of generating ground-truth SMs is discussed in Section II. The proposed algorithm is introduced in Section III. Results and performance evaluation of the proposed algorithm are shown in Section IV. Finally, conclusions and future work are provided in Section V.
II. GROUND-TRUTH GENERATION

In the psychophysical experiment in [7], several observers were asked to manually segment a set of images to generate meaningful maps. A variety of simple to complex portrait and landscape type RGB color images were used in the study. Fig. 1 shows examples of three images with their human-generated SMs. The experiment yielded several valid segmentations for the target images. If one considered any image in Fig. 1, the given SMs capture different level of details from the image and merge (ignore) others. This is mainly because image segmentation is an ill-posed problem which leads us to the definition of a segmentation spectrum model. It is the range of all possible segmentations for a given image. This includes human- and machine-generated SM. Fig. 2 portrays the segmentation spectrum and Fig. 3 shows how it is used in our algorithm.

To give a closer look of the segmentation spectrum, we utilized the segmentation algorithm in [6] (discussed later in this section) and saved the segmentation map in every iteration of the merging module. The term “segmentation spectrum” is used to describe all possible segmentation maps of a target image. The extreme over-segmentation map categorizes every pixel to independent segment. On the other extreme where the ultimate under-segmented map lands all image pixels to one region as shown in Fig. 2. A more detailed representation of the segmentation spectrum is found in Fig. 3 where several segmentation maps of the given image are shown. Some of these SMs are acceptable and the others are impractical. The acceptable SMs are either human-generated or machine-generated while the impractical SMs are all machine-generated. In Fig. 3 and Fig. 4, human observers decided about the acceptable and impractical machine-generated SMs (The study that shows this classification is discussed later in this section). We have included this representation of segmentation spectrum to illustrate the proposed idea of our segmentation evaluation algorithm.

The algorithm has been developed using the Berkeley segmentation dataset [7]. This image set contains 300 images along with their human-generated segmentations (1633 ground-truth maps) which are considered acceptable and ideal SMs. They have been used as the ideal ground-truth.

In our approach, we proposed to develop an unsupervised (stand-alone) algorithm that is capable of estimating the usefulness of a given segmentation. Thus, using the ideal segmentation for training the proposed algorithm is not realistic because the ideal segmentation is not available for any test image. Thus, we propose to use a range of acceptable and machine-generated segmentation as a training set in our methodology.

To this effect, the images in Berkeley dataset are segmented using the algorithm in [6] to generate several partition maps with various numbers of segments. The technique in [6] is an unsupervised color image segmentation algorithm. It is a gradient-based technique that uses dynamic generation of clusters to generate an initial segmentation map. Furthermore, it fuses color information (in CIELAB color space) and texture models (using local entropy) to group pixels with similar characteristics. A region growth phase followed by a unique iterative multi-resolution merging procedure is used to develop the final segmentation map of the target color image. The segmentation map in each merging iteration is saved and used as a realization in the segmentation spectrum as shown in Fig. 2.

The segmentation maps and the original images were used to develop the ground truth maps where two human observers manually selected equal number (up to four) of “acceptable” and “impractical” maps for each image.

Fig. 1. Ideal Segmentations generated by human observers from the Berkeley segmentation dataset [7].

Fig. 2. Segmentation Spectrum
The restriction of equal number of acceptable and impractical SMs has been employed to have unbiased training database and thus have equal prior probability of both classes in the BN.

An example of the ground-truth maps (machine-generated) is shown in Fig. 4 where the set of acceptable segmentations contains semantically meaningful partition maps. That is, they are visually consistent with the original image and have the right number of homogenous segments. On the other hand, the impractical segmentations are the ones that have either over partition image regions or merge semantically different regions into one segment.

Of the 300 images in the Berkeley database, a set of 150 images (randomly selected) are used to structure and train the BN while the other 150 images (test dataset that are not included in the training set) are used for evaluating the proposed algorithm as discussed in Sections III and IV.

### III. PROPOSED ALGORITHM

The proposed framework (illustrated in Fig. 5) is an image segmentation evaluation system that utilizes BN technology to estimate the probability of usefulness of a segmentation map. The proposed methodology has two phases, supervised learning phase and unsupervised testing phase. The former is to discover the optimum network structure and to estimate the conditional probabilities of its nodes. The later is to utilize the learned knowledge about the network to estimate how meaningful a segmentation of a test image is in terms of its probability. Modules that are included in the proposed methodology are discussed in this section.

#### A. Feature Extraction

Several global and local image features are used in a probabilistic framework to distinguish between acceptable and impractical segmentation maps in the proposed algorithm. They are summarized as follows:

**Feature 1:** It is the standard deviation of the normalized area of the image segments weighted by the number of segments. Feature 1 is defined as follows:

\[
F_1 = N \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 = \sqrt{N \sum_{i=1}^{N} (x_i - \mu)^2} \tag{1}
\]

where \(N\) is the number of segments, \(x\) is the normalized segment area and \(\mu\) is the mean value of the normalized segment areas. This feature is utilized to balance the number of segments and their relative size in the target segmentation map. This feature has been developed to mimic the behavior of the human observers. Most of the ideal SMs found in the Berkeley database fall towards the under-segmented side of the spectrum. This is due to the fact that human pays more attention to the object (semantic) level but not to the small details which yields in segmentation with coherent segment size (low segment-size variance). Notice that this feature does not depend on the image content but rather on the number and relative size of the segments.

**Feature 2:** It is the entropy of the lightness channel \((L')\) in the CIELAB color space weighted by the number of regions of the SM. It is defined as follows:

\[
F_2 = NH = -N \sum_{x=0}^{M_1-1} \sum_{y=0}^{M_2-1} P(x,y) \log_2 P(x,y) \tag{2}
\]

where \(N\) is the number of segments, \(H\) is the image entropy which is a statistical measure of randomness. \(P(x,y)\) in (2) is the lightness value for an input image with size \(M_1 \times M_2\). Feature 2 relates the image content to the optimal number of segments that would model it.
**Feature 3:** This attribute captures the error in the color gradient channel caused by the test SM. The error values of different color gradient regions are modeled as independent and normally distributed Random Variables (RV). The weighted sum of these Gaussian RVs (color gradient error values) is used in the proposed algorithm as Features 3. The optimal segmentation map should partition the target image into uniform “homogenous” gradient regions. In order to model that, the following steps were utilized: 1) The color gradient $g$ of the RGB color image is computed for every pixel using the algorithm in [8]. 2) The test segmentation map is superimposed over the target color gradient error map. 3) In each segment, the mean gradient value is calculated to generate $e_i = g(x,y) - \hat{g}$ where $\hat{g}$ is the average value. 4) The squared error value $e_i^2$ is found to generate a color gradient error map. 5) The weighted sums of mean values of the color gradient error map are calculated as following:

$$F_3 = \sum_{i=1}^{N} e_i^2 a_i$$

(3)

Where $a_i$ is the normalized segment area, and $N$ is the number of segments in the SM.

**Feature 4:** The color information of the input image is considered where a color error metric is defined as independent and normally distributed RVs for different image regions. We try to minimize the overall error such that each segment groups pixels of relatively homogenous colors. The CIELAB color space is utilized in this module where Feature 4 is generated utilizing the following steps: 1) The RGB color image is converted to CIELAB color space. 2) The test segmentation map is superimposed over the target color image. 3) In each segment, the mean values of $L^*$, $a^*$ and $b^*$ channels are computed and sorted into a vector. 4) In each segment, the second norm ($L_2$) value between the mean vector of the segment and the CIELAB values [$L^*, a^*, b^*$] at each pixel is computed. This generates a grayscale error map. Finally, 5) The weighted sum of mean values the error map is calculated as in (3) where $e_i$ is the mean error value of each segment.

Four local and global low-level features have been used in the proposed algorithm. The error in the RGB color channels has been used in the literature of image segmentation evaluation before as found in survey by Zhang et al. [3]. However, we used the CIELAB color space because of its preserved linearity to the human visual system to compute the weighted error in the Bayesian framework. In addition, utilizing the same concept in the color gradient is another contribution in our work. It is worth noticing that these error values are minimized (zero value) when each pixel is considered as independent segment. This yields in fewer over-segmented maps.

**B. Discretization**

The features generated are continuously valued maps. A discretization step based on equal frequency binning is utilized as shown in [9]. It employs a simple unsupervised and univariate discretization methodology that discretizes the continuous valued attributes based on a specified number of bins. The equal-frequency discretization method potentially suffers much attribute information loss since number of intervals is determined without reference to the properties of the training data. Therefore, the wrapper based methods [10] is utilized to overcome this drawback by refining the discretization of the continuous explanatory attributes by taking feedback from the algorithm outcome. Thorough experimental results suggest that quantizing the continuous features to two discrete regions provides the best performance in identifying the acceptable segmentation maps for our test image set. The equal-frequency discretization process is essential to simplify the computational complexity of the BN.

---

**Fig. 5.** Block diagram of the proposed algorithm; (a) supervised learning phase, and (b) unsupervised testing phase.
C. Ranking Module

A Bayesian network is used in the proposed algorithm to evaluate segmentation maps according to their usefulness in computer vision and image processing applications. A training set of 670 ground-truth segmentation maps (corresponding to 150 color images) were used to extract local and global image features. In our algorithm, the parameters are fully available but the structure is unknown. We found that our approach generated best results using the assumption that the given features are conditionally independent given the class (acceptable/impractical) of the SM in the training database. This structure corresponds to the naive BN as shown in Fig. 6 where F1, F2… F4 indicate the features in use and C is the decision node.

IV. RESULTS AND DISCUSSIONS

Image segmentation ranking generated by the proposed algorithm is shown in Fig. 7. The number of segmentations varies depending on the image content. That is, images capturing complex scenery have more SMs than simple images. The proposed algorithm categorizes the test SMs into four levels of usefulness. The tag-number on the SM in Fig. 7 indicates these categories. Category 1 shows that the results correlate well with human expectation; categories 2 and 3 illustrate that the SMs matched fairly and weakly, with human-generated SMs, respectively. Finally, category 4 indicates that the result does not correlate well with human expectation.

Notice that the proposed algorithm ranks SMs with acceptable number of segments to the highest level (category 1). These SMs capture acceptable level of details in the test image as shown in Fig. 7(a), 7(d), and 7(e). More detailed SMs have been also classified as category 1 as shown in Fig. 7(b), 7(e), and 7(f). Notice that Fig. 7(b) has more than one SM at the first level of priority with various details. Furthermore, the proposed algorithm classifies under-segmented maps to category 2 (second level of acceptable segmentations). These SMs show fewer details, however, they are acceptable for foreground-background classification and region of visual interest detection algorithms. Finally, categories 3 and 4 show the least favorable SMs where extreme under-segmented and over-segmented maps are placed. Sometimes it is a challenge to generate an acceptable machine-generated SM due to the complexity of the target image. Difficulty occurs where there is a lack of color or lightness contrast in the image as it is the case in Fig. 7(f).

Although a stand-alone system for segmentation evaluation has been proposed in this paper, we still need human-generated SM to evaluate the system’s performance. To this effect, the Berkeley segmentation dataset [7] has been used in addition to the Normalized Probabilistic Rand (NPR) index [11]. The NPR metric is an objective metric that compares a machine-generated SM to a set of human-generated SM. It has been used to evaluate the segmentation algorithm in [6]. The values of the NPR range from -1 to 1 where -1 indicates no match and 1 is an indication of a perfect match of the machine-generated SM to all human-generated SMs. This is not physically possible (one machine-generated SM cannot match more than one human-generated SM), thus the NPR index never have the value of 1 [11]. In this paper, we have modified the NPR to evaluate several machine-generated SMs to several human-generated SMs. This is done simply by averaging the NPR values of the same image.

In Section II, the methodology for generating the ground-truth maps used in training and testing the proposed algorithm has been described. The test set contains 150 images with their corresponding SMs. It is worth noticing that of the 1007 SMs in the test database, a set of 670 SMs are categorized to acceptable and impractical sets by human observers (as shown in Table 1). The other 337 SMs are undecided. Human observers do not have enough confidence to classify these SMs to acceptable or impractical maps. The undecided SMs usually show acceptable segments at one region but under-segment (or over-segment) other regions. The NPR average values for the test image-set (150 image) are shown in Table 1. The values are 0.4905 for the 335 acceptable segmentations and 0.1941 for the 335 impractical SMs. These NPR indexes serve as the upper and lower bounds for the test set.

Table 2 quantifies the results of the proposed algorithm where the SMs at the first level (category 1) have an average NPR value of 0.4370 based on 409 SMs. The NPR value of the other categories has decreased as shown in Table 2. This is similar to the case of percentage of SMs with NPR > 0.7 where SMs in category 1 have the highest percentage.

V. CONCLUSIONS AND FUTURE WORK

An image understanding algorithm for identifying acceptable segmentation maps has been introduced. It is an unsupervised (stand-alone) methodology based on Naïve Bayesian Networks. The algorithm has modules for image segmentation, feature extraction, discretization, and probabilistic modeling. Images from the Berkeley segmentation dataset has been used to train and evaluate the proposed algorithm. The algorithm ranks several machine-generated segmentations in accordance to their usefulness into four categories. Performance of the proposed algorithm has been evaluated using the Normalized Probabilistic Rand index. Future work includes expanding the test database and exploring other spatial and spectral features to enhance the system’s performance. Applications of the proposed algorithm include content-based image retrieval, adaptive image compression and coding, and automatic image annotation.
Fig. 7. Results from the proposed image segmentation evaluation algorithm.

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Number of SMs</th>
<th>NPR Average Value</th>
<th>NPR Standard Deviation</th>
<th>SMs with NPR &gt; 0.7 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptable SM</td>
<td>335</td>
<td>0.4905</td>
<td>0.3221</td>
<td>29.2537</td>
</tr>
<tr>
<td>Impractical SM</td>
<td>335</td>
<td>0.1941</td>
<td>0.5529</td>
<td>14.0299</td>
</tr>
</tbody>
</table>

Table II: NPR values for test image set (machine-generated SMs)

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of SMs</th>
<th>NPR Average Value</th>
<th>NPR Standard Deviation</th>
<th>SMs with NPR &gt; 0.7 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>307</td>
<td>0.4247</td>
<td>0.4301</td>
<td>27.6873</td>
</tr>
<tr>
<td>2</td>
<td>410</td>
<td>0.3788</td>
<td>0.4602</td>
<td>22.1951</td>
</tr>
<tr>
<td>3</td>
<td>221</td>
<td>0.3672</td>
<td>0.4391</td>
<td>22.1719</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>0.3259</td>
<td>0.4154</td>
<td>15.2174</td>
</tr>
</tbody>
</table>

REFERENCES
