Chapter 9

Evaluation of the segmentation methods

This chapter concentrates on evaluation of the results of segmentation performance for the texture gradient watershed method and the two other segmentation methods implemented in our work. Our evaluation measure is mainly related to the consistency between segmentations. We use segmentation error measures that provide an objective analysis of the segmentation algorithms. The evaluation parameters are Global Consistency Error and Local Consistency Error. For verifying the consistency of the evaluation results, quantitative evaluation is implemented using the measures of precision and recall. The main objective of this part of our research is to offer quantitative evidence that the use of Texture Gradient Watershed method results in superior segmentation quality.

9.1 Introduction

Despite significant advances in image segmentation techniques, evaluation of these techniques thus far is largely subjective. Segmentation is a frequent pre-processing step in many image understanding algorithms and practical vision systems [58]. In an effort to compare the performance of current segmentation algorithms to human perceptual grouping as well as understand the cognitive processes that govern grouping of visual elements in images, much work has gone into amassing hand-labeled segmentations of natural images [102]. Quantifying the performance of a segmentation algorithm, however, remains a challenging task. This is largely due to image segmentation being an ill-defined problem – there is no single ground truth segmentation against which the output of an algorithm may be compared. Rather the comparison is to be made against the set of all possible perceptually consistent interpretations of the image, of which only a minuscule fraction is usually available [58]. However an elaborate study in this area is unfeasible. Therefore this comparison is made by quantifying the agreement of output segmentation with the inherent variation in a set of available manual segmentations.
For this purpose, the Corel Segmentation Database with human segmented images [28] is used in the experiments. In this chapter, the problem of evaluating the segmentation quality of the three methods is examined with suitable metrics that are computed efficiently. Finally, we present a quantitative comparison of the results of segmentation obtained against these algorithms on the test images.

9.2 Segmentation Evaluation

Image segmentation and recognition are central problems of image processing for which we do not yet have any general purpose solution approaching human-level competence. Recognition is basically a classification task and an empirical estimate of the recognition performance (probability of misclassification) is obtained by counting classification errors on a test set [58]. Today, reporting recognition performance on large data sets is a well-accepted standard. In contrast, segmentation performance evaluation remains subjective [145]. Typically, results on a few images are shown and the authors argue why they look good. The readers frequently do not know whether the results are opportunistically selected or are typical examples, and how well the demonstrated performance extrapolates to larger sets of images.

The main challenge is that the question “to what extent is this segmentation correct” is much more subtle than “is this face from person x”. While a huge number of segmentation algorithms are reported, there is only little work exposed on methodologies of segmentation performance evaluation [146]. Several segmentation tasks are identified: edge detection, region segmentation, and detection of curvilinear structures. Their performance evaluation is of quite different nature. For instance, an evaluation of detection algorithms for curvilinear structures must take the elongated shape of this particular feature into account.

Evaluation methods are broadly divided into two categories: analytical methods and empirical methods [168]. The analytical methods directly examine and assess the segmentation algorithms themselves by analyzing their principles and properties [168]. The empirical methods indirectly judge the segmentation algorithms by applying them to test images and measuring the quality of segmentation results [168]. The former method does not suffer from influences caused by the arrangement of evaluation experiments as the latter. Although using analytical methods avoids the implementation of these algorithms, they are not much attended mainly because of the difficulty to compare algorithms solely by analytical studies.
Empirical methods are further classified into two types: *goodness* methods and *discrepancy* methods [168]. In the empirical goodness methods some desirable properties of segmented images, often established according to human intuition, about what conditions should be satisfied by an “ideal segmentation,” are measured by goodness parameters [168]. The performance of the segmentation algorithms under study is judged by the values of goodness measures [58][146]. Different types of goodness measures as entropy, color uniformity, region shape etc., are proposed in the literature [160][168]. Empirical discrepancy methods are based on the availability of *reference segmentation*, also called *gold standard* or *ground truth*. The disparity between an actually segmented image and a correctly/ideally segmented image (the gold standard, which is the best expected result) is used to assess the algorithm’s performance [168]. Both images (actually segmented and reference) are obtained from the same input image. The methods in this group take the difference (measured by various discrepancy parameters) between the actually segmented image and the reference one into account, i.e., these methods try to determine how far the actually segmented image is from the reference image.

It is hard to establish the evaluation measure for segmentation by considering various kind of performance metrics required to meet the objective of the segmentation. However, generally, segmentation performance is evaluated based on three types of metrics namely *accuracy*, *precision* and *efficiency* in order to avoid error in results [58].

- **Accuracy**: a measure of how well the segmentation output agrees with human perception.
- **Efficiency**: a measure of amount of time or effort required to perform segmentation.
- **Precision**: a measure of degree of the same result produced over different segmentation sessions.

A potential problem for a measure of consistency between segmentations is that there is no unique segmentation of an image [58][61]. For example, two people segment an image differently because either they perceive the scene differently, or they segment at different granularities. If two different segmentations arise from different perceptual organizations of the scene, then it is fair to declare the segmentations inconsistent. If, however, one segmentation is simply a refinement of the other, then the error is to be small, or even zero as illustrated in Figure 9.1.

In addition to being tolerant to refinement, any error measure is also

1. independent of the coarseness of pixelation
2. robust to noise along region boundaries, and
3. tolerant of different segment counts between the two segmentations.

The third point is due to the complexity of the images. The requirement is the comparison between segmentations when they have different numbers of segments.

![Fig. 9.1: (a) original image (b) - (d) three different segmentations](image)

**9.3 Error Measures for validation**

Segmentation is simply a division of the pixels of an image into sets. A segmentation error measure takes two segmentations $S_1$ and $S_2$ as input, and produces a real valued output in the range $0 \ldots 1$ where zero signifies no error [58][146]. The measures are designed to be tolerant to refinement, that is, if subsets of regions in one segmentation consistently merge into some region in the other segmentation, the consistency error is low [58] [145]. For a given pixel $p$, consider the segments in $S_1$ and $S_2$ that contain that pixel. The segments are sets of pixels. If one segment
is a proper subset of the other, then the pixel lies in an area of refinement, and the local error is zero \[61][145\]. If there is no subset relationship, then the two regions overlap in an inconsistent manner. In this case, the local error is non-zero \[61][145\]. Let \(\setdiff\) denote set difference, and \(|x|\) the cardinality of set \(x\). If \(R(S, p_i)\) is the set of pixels corresponding to the region in segmentation \(S\) that contains pixel \(p_i\), the local refinement error is defined as:

\[
E(S_1, S_2, p_i) = \frac{|(R(S_1, p_i) \setdiff (R(S_2, p_i))|}{|R(S_1, p_i)|}
\tag{9.1}
\]

Note that this local error measure is not symmetric. It encodes a measure of refinement in one direction only: \(E(S_1, S_2, p_i)\) is zero precisely when \(S_1\) is a refinement of \(S_2\) at pixel \(p_i\), but not vice versa. So for every pixel it is computed twice, once in each direction. Given the error measures at each pixel, two segmentation error measures are defined for the entire image, Global Consistency Error (GCE) and Local Consistency Error (LCE) \[61][145\]. Let \(n\) be the number of pixels: Then

\[
GCE(S_1, S_2) = \frac{1}{n} \min \left( \sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i) \right)
\tag{9.2}
\]

\[
LCE(S_1, S_2) = \frac{1}{n} \sum \min \left( E(S_1, S_2, p_i), E(S_2, S_1, p_i) \right)
\tag{9.3}
\]

The GCE assumes that one of the segmentations must be a refinement of the other, and forces all local refinements to be in the same direction. The LCE allows for refinements to occur in either direction at different locations in the segmentation. It is clear that GCE is a tougher measure than LCE. Considering Figure 9.1, GCE tolerates the simple refinement from (c) to (b) or (d), while LCE tolerates the mutual refinement of (b) and (d). Note that since both measures are tolerant of refinement, they are meaningful only when comparing two segmentations with an approximately equal number of segments. This is because there are two trivial segmentations that achieve zero error: One pixel per segment, and one segment for the entire image. The former is a refinement of any segmentation, and any segmentation is a refinement of the latter.

Images from the Corel Dataset \[28\] for which manual segmentations or ground truth images provided by experts are available are used for evaluations. This database of 40,000 images is widely used in computer vision. Around 1000 representative images from this image database is
chosen to form the evaluation set. The criterion for selecting images is based on images that contain at least one discernible object and with rich texture content.

Fig. 9.2: (a) Original images from the Corel dataset (b) – (d) different ground truths
The human segmentations though varying in detail are consistent with one another in that regions segmented by one subject at a finer level of detail can be merged consistently to yield the regions extracted by a different subject at a coarser level of detail as shown in Figure 9.2.

The GCE and LCE measures are computed for all the segmentations in our dataset with two goals. First, we hope to show that given the arguably ambiguous task of segmenting an image into an unspecified number of segments, different people produce consistent results on each image. Second, we hope to validate the measures by showing that the error between segmentations of the same image is low, while the error between segmentations of different images is high.

Each texture feature extraction technique requires a set of parameters to characterize textures. In the experiments, in the segmentation with GLCM and Bayesian classifier, the parameters are constituted by the orientation angles of 0°, 45°, 90° and 135°, spatial resolution of 1, 3 and 5 and sampling window size of 7 and 11.

In the second case of Texture gradient watershed method (TGWS), the decomposition scaled from 2 to 5 for the various test images thereby making the feature images number extend from 12 to 336. The minsize parameter varies from 500 to 2000 for the images for effective segmentation.

In the third case of Gabor filters and SOM, the parameters are constituted by the orientation angles of 0°, 45°, 90° and 135° and five radial frequencies.

A histogram-based evaluation mechanism is elaborated, aimed at comparing the segmentation results for the experimented algorithms via the errors metrics. In order to achieve this, we consider the provided ground-truth segmentations and compare the segmentations of each algorithm with it, measuring the error metrics GCE and LCE.

For every selected image, four ground truth images from four experts are considered. The GCE and LCE are computed in each case of comparing the segmented image against each ground truth. The average of the error metrics obtained is taken as the input for generating the histograms. The histograms presented in Figures 9.3 to 9.10 illustrate this approach.

For a better description of the histogram based analysis, in Figures 9.3 and 9.4, the distribution of the values of GCE and LCE respectively are shown for the images processed using TGWS algorithm.
Fig. 9.3: GCE for Texture Gradient Watershed method

Fig. 9.4: LCE for Texture Gradient Watershed method

Fig. 9.5: GCE for Gabor Filter and Self Organizing Map method
Fig. 9.6: LCE for Gabor Filter Self Organizing Map method

Fig. 9.7: GCE for GLCM and EM - Bayesian method

Fig. 9.8: LCE for GLCM and EM-Bayesian method
It is very important to note that these values are mostly concentrating on smaller error values, which indicates the superior performance of this method with respect to the other two. A comparison between the error metrics of all the three methods is show in Figures 9.9 and 9.10; it gives a good perspective on the error values generated by each scheme. It also shows that, when pairs of human segmentations of the same image are compared, both the GCE and the LCE are low; conversely, when random pairs of human segmentations are compared, the resulting GCE and LCE are high.

9.4 Boundary based evaluation measures
Segmentation evaluation metrics are also divided into boundary-based methods and region-based methods[48][61]. GCE and LCE measures are widely used for evaluation of region based
methods. For boundary-based evaluation, the use of precision-recall curves is common [101]. For comparison of the results obtained with the adopted error measures, the measure of precision-recall is also used in our experiments with evaluation. Recall is defined as the proportion of boundary pixels in the ground truth that are successfully detected by the automatic segmentation while precision is the proportion of boundary pixels in the automatic segmentation that correspond to boundary pixels in the ground truth [48][101]. Precision and recall are attractive as measures of segmentation quality because they are sensitive to over-segmentation and under-segmentation. Over-segmentation leads to low precision scores, while under-segmentation leads to low recall scores [48][101]. A comparison of two segmentations yields high values of both precision and recall, only if the boundaries in both segmentations agree in location and level of detail. Additionally, if the region boundaries are in agreement, the regions themselves must agree across the segmentations.

Precision and recall are estimated by computing a minimum cost matching between the boundary pixels in two segmentations [101]. The matching cost strategy is computed as follows:

Given two segmentations $S_1$ and $S_2$, we find a suitable match for each boundary pixel in $S_1$ by examining its neighborhood within a radius of $r$ for boundary pixels in $S_2$. We match a pixel $p_i$ in $S_1$ to a pixel $p_x$ in $S_2$ if

- There is no other boundary pixel $p_j$ in $S_1$ between $p_i$ and $p_x$, with the exception of $p_i$’s immediate neighbor
- The nearest boundary pixel in $S_1$ for $p_x$ is in the general direction of $p_i$ (though it does not have to be $p_i$). If $p_x$ has several nearest neighbors, at least one of them must point in the general direction of $p_i$ (in practice, this means that the directions from $p_x$ to $p_i$, and from $p_x$ to one of its nearest neighbors should be within 25 degrees of each other).

If more than one pixel in $S_2$ satisfies the above conditions for a given $p_i$, the nearest one is chosen. Together, these conditions imply that $p_i$ is not matched to a boundary pixel $p_x$, unless $p_i$ is part of the closest boundary in $S_1$ to $p_x$’s boundary in $S_2$. The direction condition is necessary to avoid double matches when a boundary in $S_2$ is flanked on both sides by boundaries in $S_1$.

Precision is now defined as

\[
\text{Precision} (S_1, S_2) = \frac{|\text{unmatched} (S_1)|}{|S_1|}
\]

where unmatched ($S_1$) is the set of boundary elements in $S_1$ that do not have a suitable match in $S_2$ within a distance $r$ and $|\cdot|$ represents set cardinality.
Similarly, recall is defined as \( \text{Recall}(S_1, S_2) = \left| \text{unmatched}(S_2) \right| / \left| S_2 \right| \)

However, instead of comparing the algorithm’s output against several different human segmentations of the same image, it is compared against the union of the boundaries from all human segmentations of that same image [102]. The composite segmentation contains any boundaries that human observers considered salient, and thus provides a more useful benchmark in testing for over and under-segmentation. Additionally, the fact that even though different observers segment an image at different levels of detail, some of the boundaries in the image, agreed upon by most (and in fact, usually by all) human observers, are also taken into consideration. Boundaries in the composite segmentation are weighted according to how many observers marked the same boundary pixel, normalized by the number of observers. This yields a composite image with values in \([0, 1]\), where 0 represents no boundary, and 1 represents a boundary pixel that was marked by all the human observers segmenting the image. This is illustrated in Figure 9.11(b).

From the above definitions, precision is low when there is significant over-segmentation, when the localization errors of significant portions of the boundary is greater than \( r \), or when a significant portion of the boundary pixels in \( S_1 \) are matched to pixels in \( S_2 \) with low boundary weight (meaning that they capture structure that was not considered salient by many human observers). Similarly a low recall value is typically the result of under-segmentation and the failure to capture salient image structure, especially if the unmatched boundaries have high boundary weight in \( S_2 \) as illustrated in Figure 9.12.
Figure 9.12 shows two segmentations namely over and under-segmentation of the composite segmentation in Figure 9.11 (b). The figure also shows the precision and recall for each case. In the case of the over-segmented image, the recall is high with low precision, indicating detection of most of the boundaries in the composite human segmentation. The under-segmented image receives low recall since it fails to capture salient image structure. The low precision in this case indicates a relatively large proportion of the boundary pixels in the segmentation remaining unmatched. An interesting observation from the results on the under-segmented image is that it is possible for algorithms to under-segment globally while over-segmenting locally.
Precision and recall measures are attractive for one additional reason [48][101]. Precision/ recall curves are used to characterize the performance of an algorithm over a range of input parameters (thus allowing for the selection of the optimal parameters for any desired recall value), and to compare different algorithms on an equal footing. For algorithms with a single input parameter, a single curve is obtained while for algorithms with two parameters, a curve for each possible value of one input parameter is obtained and the values along the curve correspond to variations of the second parameter.

The segmentations produced by two of our algorithms consist of labeled images in which pixels within the same region have an identical label. Since our metrics require region boundaries instead of region label maps, boundary detection is done in these cases before computation of precision and recall. To ensure fairness in the evaluation process, the boundaries are extracted using an identical procedure for all the three algorithms. The procedure consists of taking the labeled image output of the algorithms and marking each pixel that has at least one neighbor with a different label as a boundary pixel. This causes a 2-pixel wide boundary to be marked along region borders and wider boundaries at junctions of more than two regions.

The input parameters and ranges are as follows:
For the TGWS method the only input parameter is the minsize parameter varying from 500 to 2000.
For the second case of GLCM and EM-Bayesian classification method, there are two parameters: the orientation angles of 0°, 45°, 90° and 135°, spatial resolution of 1, 3 and 5. The sampling window size is set to 11.
In the third case of Gabor filters and SOM, there are two parameters constituted by the orientation angles of 0°, 45°, 90° and 135° and three spatial bandwidths.
The neighborhood radius r is set to 3 for all the cases.

Figure 9.13 shows the tuning curves obtained for the various cases. These curves facilitate selection of appropriate parameter values which in turn enables a target value of either precision or recall for a given algorithm (limited by the operational range of the algorithm). Most importantly, they provide a direct way of comparing the quality of the segmentations produced by different algorithms across a wide range of input parameters. It enables judging the performance of algorithms consistently better than others across its particular range of parameters and rank algorithms by performance for particular values of precision or recall. The
superior performance of the TGWS segmentation method is proven in Figure 9.13 (d). At the point of highest recall, GLCM-EM method produce notoriously over-segmented results, while at the point of highest precision Gabor-SOM method produce segmentations that capture small, high-contrast regions that do not necessarily capture the structure of the image. Over-segmentation is limited for TGWS method by the use of metadata items describing spatial arrangements for textures. Over-segmentation is characterized in the curves by high recall but low precision, and the converse is true for under-segmented images.
9.5 Summary

Meaningful comparison of image segmentation methods is often dependant on the application. For some applications (e.g. medical images) the aim of the segmentation is to identify perceptually salient image elements such as areas of tumors and high contrast regions. Evaluation of segmentation techniques for such applications is effectively based on a perceptual comparison. For other applications, computational measures are also used for segmentation evaluation. The qualitative comparisons are already presented in the preceding chapters. The aim of this chapter is to evaluate quantitatively the segmentation result obtained using the Texture gradient watershed method to the other two methods. Ground truth images of the Corel Database are used as reference images for the evaluation process. The error evaluation metrics adopted are the Local Consistency Error and Global Consistency Error measures. To establish the reliability of the adopted measures for performance assessment, an alternate approach using the precision-recall measures is also implemented for evaluation. The superior performance of the Texture Gradient Watershed method is explicit owing to the scale selectivity, shift invariance and orientation properties of the DTXWPT. The Gabor filters also perform comparatively well because of their capability of characterizing textures through the dominant frequency and orientation components. The statistical method was found to exhibit the worst performance owing to the limited spatial information and the high computational complexity.

The contributions of this chapter are:

• Reviewed the various methods of segmentation performance evaluation
• Implementation of evaluation methodology and computing the evaluation parameters of LCE and GCE as well as Precision-Recall
• Implemented histogram based evaluation aimed to compare the segmentation results for the experimented algorithms via the errors metrics.
• Demonstrated the superior performance of the Texture Gradient Watershed method over the other methods
• Formation of composite ground-truth image
• Generation of Tuning curves for individual methods as well as for the combined case