Category: Multimedia Technology

Image Segmentation Evaluation in this Century

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

Image segmentation consists of subdividing an image into its constituent parts and extracting those parts of interest (objects). Due to its importance in image analysis, many research works have been conducted for this process. After 40 years of development, a large number of image (and video) segmentation techniques have been proposed and utilized in various applications (Zhang, 2006). With many algorithms developed, some efforts have been spent also on their evaluation, and these efforts have resulted around 100 evaluation papers that can be found in literature for the last century. Several studies have been made in the past in attempt to characterize these existing evaluation methods (Zhang, 1993; Zhang, 1996; Zhang 2001).

Segmentation evaluation methods can be classified into analytical methods and empirical methods (Zhang, 1996). The analysis methods treat the algorithms for segmentation directly by examining the principle of algorithms while the empirical methods judge the segmented image (according to predefined criteria or comparing to reference image) so as to indirectly assess the performance of algorithms.

Empirical evaluation is practically more effective and usable than analysis evaluation (Zhang, 1996). Recent advancements for segmentation evaluation are mainly made by the development of empirical evaluation techniques. After providing a list of evaluation criteria and methods proposed in the last century as background, this article will provide a summary of the recent (in 21st century) research works for empirical evaluation of image segmentation. These new research works are classified into three groups: (1) those based on existing techniques, (2) those made with modifications of existing techniques, and (3) those that used dissimilar ideas than that of existing techniques. A comparison of these evaluation methods is made before going to the future trends and conclusion.

BACKGROUND

Empirical evaluation methods can be classified into *goodness method group* and *discrepancy method group* (Zhang, 1996). They use different empirical criteria for judging the performance of segmentation algorithms. The goodness method can perform the evaluation without the help of reference images while the discrepancy method needs some reference images to arbitrate the quality of segmentation. In Zhang (1996) the eight mostly used criteria (three goodness ones and five discrepancy ones) have been discussed in details. All these criteria have been grouped into a table in Zhang (2001, pp. 148-151), and the table is reproduced in Table 1.

There are other criteria discussed in Zhang (1996), though they were not very well liked in that time. One is the number of regions in a segmented image. In case no ground truth was available, it would be expected to get a modest result, so the *moderate number of regions* could be counted as a criterion. Some others come from the class D-5, such as *region consistency*, *grey level difference*, and *symmetric divergence (cross-entropy)*. Finally, several criteria used in special methods have attracted certain attention recently, such as *amount of editing operations, visual inspection,* and *correlation between original image and segmented bi-level image.* All these criteria are listed now in Table 2 as a complementary of Table 1.

MAIN FOCUS OF THE CHAPTER

Getting into the new century, the research on segmentation evaluation has attracted even more attention in the segmentation community. In the following section, some evaluation works published since 2002, that is, after the last review paper on evaluation (Zhang, 2001), are sketched and discussed. Among these new empirical evaluation works, some are based on existing techniques, some are made with modifications/improvements of existing techniques, and some have dissimilar ideas than that of existing techniques.

Evaluation Works Based on Existing Techniques

In Cavallaro, Gelasca, and Ebrahimi (2002), a single objective metric is formed by using both spatial and temporal consistency information. The metric was defined based on two types of errors. One is the number of (both positive and negative) false pixels. Another is the distance of false pixels to their correct places. The spatial context was introduced to weight the false pixels according to their distance to the

Table 1. A list of empirical criteria and their method groups

Class	Criterion name	Method group
G-1	Intra-region uniformity	Goodness
G-2	Inter-region contrast	Goodness
G-3	Region shape	Goodness
D-1	Number of mis-segmented pixels	Discrepancy
D-2	Position of mis-segmented pixels	Discrepancy
D-3	Number of objects in the image	Discrepancy
D-4	Feature values of segmented objects	Discrepancy
D-5	Miscellaneous quantities	Discrepancy

Table 2. A complementary list of empirical criteria and their method groups

Class	Criterion name	Method group
G-4	Moderate number of regions	Goodness
D-5a	Region consistency	Discrepancy
D-5b	Grey level difference	Discrepancy
D-5c	Symmetric divergence (cross-entropy)	Discrepancy
S1	Amount of editing operations	Special
S2	Visual inspection	Discrepancy like
S3	Correlation between original image and bi-level image	Goodness like

reference boundary. In addition, temporal context has been used to assign weight inversely proportional to the duration of an error for evaluating the quality variation over time. The overall metric was eventually formulated as nonlinear combination of the number of false pixels and the distances, weighted by the temporal context factor.

In Prati, Mikic, and Trivedi (2003), a comparative empirical evaluation of representative segmentation algorithms selected from four classes of techniques (two statistical ones and two deterministic ones) for detecting moving shadows has been made with a benchmark for indoor and outdoor video sequences. Two quantitative metrics: (1) good detection (low probability of misclassifying a shadow point) and (2) good discrimination (the low probability of classifying non-shadow points as shadow) are employed.

In Rosin and Ioannidis (2003), an evaluation of eight different threshold algorithms for shot change detection in a surveillance video has been made. Pixel-based evaluation is applied by using true positive (TP), true negative (TN), false positive (FP), and false negatives (FN).

In Lievers and Pilkey (2004), a comparison of 12 automatic global thresholding methods has been made. Among them, eight are point-dependent algorithms and four are region-dependent algorithms. Some multimodal images have been tested. Authors defined a cost function for selecting the appropriate thresholds. This cost function is based on intra-class variations, so it is not surprising that the best algorithm found by authors is a minimum cross-entropy method.

In Marcello, Marques, and Eugenio (2004), a survey of 36 image thresholding methods, with a view to assess their performance when applied to remote sensing images and especially in oceanographic applications, has been conducted. Those algorithms have been categorized into two groups: local and global thresholding techniques. For performance judgment, only visual inspection is carried out.

In Renno, Orwell, and Jones (2004), four different shadow suppression algorithms have been evaluated by using video from a nightly soccer match with quite some shadow because of the lighting used. All evaluation metrics are based on the number of correctly detected pixels. The metrics used are the detection rate, the false positive rate, the signal-to-noise ratio, and the tracking error (the average distance between ground truth boxes and tracked targets). Finally, using an average over time, the performances of shadow segmentation of the four shadow suppression algorithms are compared.

In Carleer, Debeir, and Wolff (2004), four algorithms were applied to high spatial resolution satellite images and their performances were compared. Two empirical discrepancy evaluation criteria are used: (1) the number of mis-segmented pixels in the segmented images compared with the visually segmented reference images, and (2) the ratio between the number of regions in the segmented image and the number of regions in the reference image.

In Ladak, Ding, and Wang (2004), a comparison of three kinds of segmentation algorithms for 3-D images: (1) segmenting parallel 2-D slice images, (2) segmenting rotated 2-D slice images, and (3) directly segmenting volume-based 3-D image, was carried out. The judging parameter used is the percent difference in volume (volume error) between automatically segmented objects and the manually determined (by a trained person) ground truth. The times needed for editing the segmented objects obtained by using the three kinds of algorithms to fit the ground truth are also compared.

Evaluation Works Made with Modifications/Improvements

In Oberti, Stringa, and Vernazza (2001), Receiver Operating Curve (ROC) is used to extract useful information about the segmentation performance when changing external parameters that describe the conditions of the scene. ROC has also been used in Niemeijer, Staal, and Ginneken (2004) for studying the performance of five vessel segmentation algorithms.

In Udupa, LeBlanc, and Schmidt (2002), three groups of factors: (1) precision, (2) accuracy, and (3) efficiency, are considered for evaluating segmentation methods in assessing