A Review of Image Segmentation Evaluation in the 21st Century

M

Yu-Jin Zhang
Department of Electronic Engineering, Tsinghua University, China

INTRODUCTION

In image engineering, computer vision and image pattern recognition, image segmentation plays an important role. It consists of subdividing an image into its constituent parts and extracting those parts of interest (objects). In the past 50 years, many research works have been conducted in this area, a large number of image (and video) segmentation techniques have been proposed and utilized in various applications. With many algorithms developed, some efforts have been spent also on their evaluation, a review for the efforts in the last century can be found in (Zhang, 2001).

The first comprehensive review on image segmentation evaluation has been made nearly 20 years ago (Zhang, 1996). The existing evaluation methods for segmentation algorithms have been classified into analytical methods and empirical methods. The analysis methods treat the algorithms for segmentation directly by examining the principle of algorithms while the empirical methods judge the segmented image to indirectly assess the performance of algorithms. Furthermore, the empirical methods can be still classified into empirical goodness methods and empirical discrepancy methods. The empirical goodness methods judge the segmentation results according to some predefined (goodness) criteria while the empirical discrepancy methods determine the quality of segmented images by comparing to some reference images.

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. In this article, after providing a list of evaluation criteria and methods proposed in the last century as background, a review of the research works made in this century (till now) for empirical evaluation of image segmentation will be provided. These new techniques are classified, comparing to the last century-developed techniques, into 3 groups: those based on existing techniques, those made with modifications of existing techniques, and those 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

As mentioned above, most empirical evaluation methods can be classified into goodness method group and discrepancy method group (Zhang 1996). 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. More importantly, they use different empirical criteria for judging the performance of segmentation algorithms. These criteria play some critical roles in determining the generality, usability, sensitivity, effectiveness and efficiency of these evaluation procedures.

In Table 1, the already reviewed and compared empirical criteria for image segmentation evaluation are summarized (Zhang, 1996; Zhang, 2001). In Table 1, three groups of criteria can be distinguished: G for goodness criteria, D for discrepancy criteria and S for specialized criteria. The criteria for the last group have some particularity so as to be different from either goodness criteria or discrepancy criteria, but the methods using these criteria can still be classified as goodness like or discrepancy like ones.

MAIN FOCUS OF THE ARTICLE

Getting into the new century, the research on image segmentation evaluation has attracted more atten-

DOI: 10.4018/978-1-4666-5888-2.ch579

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

Criterion Group	No.	Criterion Name	Method Class
Goodness Criteria	G-1	Intra-region uniformity	Goodness
	G-2	Inter-region contrast	Goodness
	G-3	Region shape	Goodness
	G-4	Moderate number of regions	Goodness
Discrepancy Criteria	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 object quantities	Discrepancy
	D-6	Region consistency	Discrepancy
	D-7	Grey level difference	Discrepancy
	D-8	Symmetric divergence (cross-entropy)	Discrepancy
Specialized Criteria	S-1	Amount of editing operations	Discrepancy like
	S-2	Visual inspection	Discrepancy like
	S-3	Correlation between original image and bi-level image	Goodness like

tion in even large community. In the following, most reprehensive empirical evaluation works published in the 21st century are briefly reviewed. These works are gathered into three groups according to their relations with respect to existing works in the last century: (1) based on existing techniques; (2) made with modifications/improvements of existing techniques; (3) have dissimilar/new principles regarding to existing techniques. In addition, their evaluation criteria are classified and the novelties are pointed.

Evaluation Works Based on Existing Techniques

In Cavallaro (2002), an objective metric is formed by using both spatial and temporal consistency information. It was defined based on two types of errors: the number of false pixels and 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 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 formulated as nonlinear combination of the number of false pixels and the distances, weighted by the temporal context factor.

In Prati (2003), a comparative empirical evaluation of representative segmentation algorithms for detecting moving shadows has been made with a benchmark for indoor and outdoor video sequences. Two quantitative metrics: good detection (low probability of misclassifying a shadow point) and good discrimination (the low probability of classifying non-shadow points as shadow) are employed.

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

In Ladak (2004), a comparison of three kinds of segmentation algorithms for 3-D images: segmenting parallel 2-D slice images, segmenting rotated 2-D slice images and directly segmenting volume-based 3-D image, was carried out. The judging parameter used is the percent difference in volume (volume er-