Image and Vision Computing 27 (2009) 1223–1227

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Short communication

An evaluation metric for image segmentation of multiple objects ☆

Mark Polak, Hong Zhang *, Minghong Pi

Centre for Intelligent Mining Systems, University of Alberta, Edmonton, Alta., Canada T6G 2E8

ARTICLE INFO

Article history:
Received 7 August 2006
Received in revised form 14 July 2008
Accepted 18 September 2008

Keywords:
Image segmentation
Evaluation
Error measure

ABSTRACT

It is important to be able to evaluate the performance of image segmentation algorithms objectively. In this paper, we define a new error measure which quantifies the performance of an image segmentation algorithm for identifying multiple objects in an image. This error measure is based on object-by-object comparisons of a segmented image and a ground-truth (reference) image. It takes into account the size, shape, and position of each object. Compared to existing error measures, our proposed error measure works at the object level, and is sensitive to both over-segmentation and under-segmentation. Hence, it can serve as a useful tool for comparing image segmentation algorithms and for tuning the parameters of a segmentation algorithm.

1. Introduction

Image segmentation is one of the basic problems in image analysis. Although extensive efforts have been made to develop image segmentation algorithms, much less attention has been paid to evaluating the performance of image segmentation algorithms. In general, evaluation methods for image segmentation can be classified into analytical and empirical evaluation methods [2]. Empirical methods, in turn, can be classified into empirical goodness methods and empirical discrepancy methods.

The analytical methods of evaluation typically focus on analyzing the properties of a segmentation algorithm, such as its processing strategy, complexity, and efficiency. The evaluation is from a theoretical point of view and does not require the actual implementation of the algorithms. However, they work only for particular models or are concerned with certain desirable properties of the algorithm. The empirical goodness methods of evaluation use the original image and the resulting segmented image. Goodness can be expressed in terms of a statistical measure such as the uniformity within segmented regions [3], inter-region contrast [9], or region shape [10]. However, without a reference image, the goodness may not be objective.

The empirical discrepancy methods, to which our proposed metric belongs, explicitly calculate the error between the segmented image and a reference (ground-truth) image. The reference image is often obtained manually with the help of a human expert, and the segmented image is from a segmentation algorithm. Common error measures are the number of mis-segmented pixels [4,5], the position of mis-segmented pixels [6], the number of objects in the image [7,14], or the geometric features of segmented objects such as area, perimeter, or sphericity [8]. Almost all empirical methods are constructed by considering image segmentation as a process of pixel labeling, except for [14] which focuses on the number of objects exclusively with regard to the sizes. Consequently, they are not appropriate for object-level evaluation.

In contrast, Martin et al. [11] proposed an interesting empirical discrepancy measure for evaluating segmentation. It is an object-by-object error measure, and is very useful to quantify the consistency between segmentations manually performed by different people of the same image who view the image at different granularities or scales. Unfortunately, it is inappropriate for segmentation applications in which the details of the segmentation in terms of the exact object boundaries are important. As an example, when an over-segmented image is simply a refined version of an under-segmented image, the Martin error measure would consider the two to be consistent and therefore correct with respect to each other.

This lack of penalty for over or under-segmentation was recognized in [13], which proposed an error measure based on the concept of partition distance. Partition distance counts the number of pixels, normalized with respect to the image size, that must be removed from the interpretation, i.e., segmentation of an image until the induced segmentation agrees with the reference image. Consequently, similar to [11], partition distance considers all levels of over and under refinements to be equally incorrect. In addition, the calculation of partition distance does not weigh objects according to their sizes. However, both of the above properties are important in many applications.

In this paper, we propose a new empirical discrepancy error measure, called object-level consistency error (OCE), which quan-

☆ This research is supported in part by NSERC, iCORE, Syncrude Canada Ltd., Matrikon, and the University of Alberta.

* Corresponding author. Address: Department of Computer Science, University of Alberta, 221 Athabasca Hall, Alta., Canada T6G 2E8. Tel.: +1 780 492 7188.
E-mail addresses: mpolak@cs.ualberta.ca (M. Polak), zhang@cs.ualberta.ca (H. Zhang), minghong@cs.ualberta.ca (M. Pi).

0262-8856/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2008.09.008

tifies the similarity (or discrepancy) between a segmented image and the ground truth image at the object level. The key novelty of the error measure is that it takes into account the existence, size, position, and shape of each fragment and penalizes both over-segmentation and under-segmentation. At the same time, it retains the properties of being normalized ($0 \leqslant OCE(I_g, I_s) \leqslant 1$ and $OCE(I_g, I_s) = 0$ only if $I_g = I_s$), symmetric ($OCE(I_g, I_s) = OCE(I_s, I_g)$), and scale invariant ($OCE(I_g, I_s) = OCE(I_s^{scaled}, I_g^{scaled})$), where I_g and I_s are the segmented and the ground-truth images, and I_g^{scaled} and I_s^{scaled} are their scaled versions, respectively. We argue that our proposed OCE can effectively serve as an objective evaluation of image segmentation algorithms and for tuning the parameters of a segmentation algorithm.

The rest of the paper is organized as follows. Section 2 describes the error measure proposed by Martin et al. Section 3 describes our proposed performance metric. The experimental results are provided in Section 4, followed by the conclusions in Section 5.

2. Martin error measure

Martin et al. [11] proposed an interesting error measure, which takes two images I_g and I_s as input, and produces a real-valued output in the range of $[0,1]$ where 0 signifies no error and 1 worst segmentation. The measure is shown to be effective for qualitative similarity comparison between segmentations by humans, who often produce results with varying degrees of perceived details, which are all intuitively reasonable and therefore "correct". On the other hand, the Martin error measure is sensitive to qualitatively different segmentations.

Assume $I_g = \{A_1, A_2, \ldots, A_M\}$ is a reference image where A_j is the jth fragment in I_g. Assume further that $I_s = \{B_1, B_2, \ldots, B_N\}$ is the segmented image where B_i is the ith fragment in I_s. Let $|A|$ represent the number of pixels in A. Martin et al. [11] define the error between fragment A_j and B_i (in a different but equivalent form) as

$$P_{ji} = \frac{|A_j \setminus B_i|}{|A_j|} \times |A_j \bigcap B_i| = \left(1 - \frac{|A_j \bigcap B_i|}{|A_j|}\right) \times |A_j \bigcap B_i|, \quad (1)$$

where $\setminus$ denotes the set difference operation and $\bigcap$ denotes the intersection. Similarly, the error between fragment B_i and A_j is defined as

$$Q_{ji} = \frac{|B_i \setminus A_j|}{|B_i|} \times |A_j \bigcap B_i| = \left(1 - \frac{|A_j \bigcap B_i|}{|B_i|}\right) \times |A_j \bigcap B_i|. \quad (2)$$

The total area of intersection between I_g and I_s is calculated by

$$n = \sum_{j=1}^{M} \sum_{i=1}^{N} |A_j \bigcap B_i|. \quad (3)$$

There are two variants of the Martin error measure, global consistency error (GCE) and local consistency error (LCE). Specifically,

$$GCE(I_g, I_s) = \frac{1}{n} \min \left\{ \sum_{j=1}^{M} \sum_{i=1}^{N} P_{ji}, \sum_{j=1}^{M} \sum_{i=1}^{N} Q_{ji} \right\}, \quad (4)$$

$$LCE(I_g, I_s) = \frac{1}{n} \sum_{j=1}^{M} \sum_{i=1}^{N} \min(P_{ji}, Q_{ji}). \quad (5)$$

Although these error metrics are calculated by grouping pixels into objects first, they unfortunately tolerate over-segmentation and under-segmentation, as a consequence of their intended purpose for comparing human segmentations. As an example, take Fig. 1 which shows a ground truth image of a single object (I_0) and three possible hypothetical segmentation results (I_1, I_2, and I_3) with varying degrees of over-segmentation. Comparing A_1 and B_1, $M = N = 1$, and $n = |A_1 \bigcap B_1|$. Since A_1 and B_1 are identical, $P_{11} = Q_{11} = 0$ and $GCE = LCE = 0$, as expected of a reasonable error measure. However, for the segmentation in Fig. 1(c), $I_2 = \{C_1, C_2\}$ where C_1 and C_2 are each one half of A_1 (assuming object boundaries are of zero-pixel width) and $M = 1$ and $N = 2$, so that $P_{1i} = \frac{1}{2}|A_1 \bigcap C_i| > 0$ and $Q_{1i} = 0$ ($i = 1, 2$) and therefore $GCE = LCE = 0$, incorrectly indicating no error. Similarly, in Fig. 1(d), $I_3 = \{D_1, D_2, D_3\}$ where D_1 is a half of A_1 and D_2 and D_3 are a quarter of A_1, respectively, so that $Q_{1i} = 0$ ($i = 1, 2, 3$) and $GCE = LCE = 0$. As can be seen, the object in Fig. 1(a) can be indefinitely over-segmented, and yet error measures (4) and (5) are insensitive to the extent of over-segmentation. The

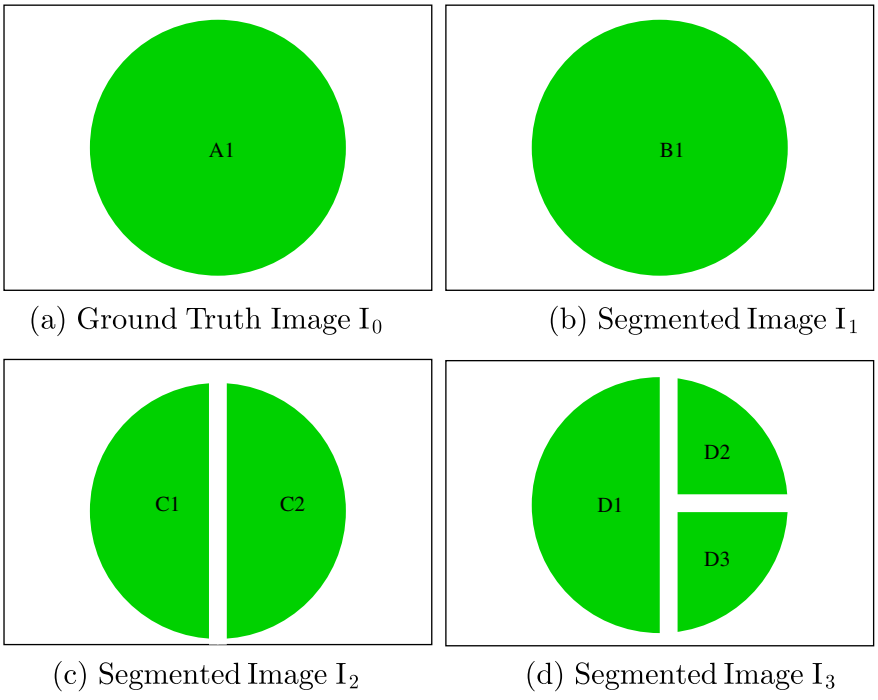

Fig. 1. The ground-truth image and three possible segmentations yield identical error scores according to Eqs. (4) and (5).