Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **6**, November **2018**

Analysis of Evaluation Metrics for Image Segmentation

Chaur-Heh Hsieh

Information Engineering College
Yango University
Fuzhou 350015, China
chaoho1204@qq.com

Tsorng-Lin Chia*

Department of Computer and Communication Engineering
Ming Chuan University
Taoyuan, Taiwan
*Corresponding author: tlchia@mail.mcu.edu.tw

Received May, 2018; Revised October 2018

Abstract. *Objective and quantitative evaluation for segmentation performance is important for the development of image segmentation algorithms. Several objective evaluation metrics have been proposed in the literature. This paper presents an analysis of the existing pixel-based and object-based evaluation metrics. We define and describe the possible error types of image segmentation. We investigate the properties of the error metrics by mathematical proof and experimental justification. The results indicate that the object-based metrics have several shortages although they are more suitable than the pixel-based metric for object-level evaluation.*

Keywords: Evaluation metrics, Image segmentation, Object-based metric, Pixel-based metric

1. **Introduction.** Image segmentation is one of the basic tasks in image and video analysis [1], [2]. Extensive efforts have been made to develop segmentation techniques, but much less attention has been paid for the performance evaluation of those techniques [3], [4]. In general, objective evaluation methods for image segmentation can be categorized into analytical methods and empirical methods. The empirical methods are further divided into discrepancy (supervised evaluation) and goodness (unsupervised evaluation) based on whether the method requires a reference (ground-truth) image or not [5]-[8]. Among them, the supervised evaluation approach is the most popular one. Generally, the performance evaluation in the supervised evaluation metric is to calculate the error between the segmented image and a reference image. The reference image is often obtained manually, and the segmented image is from a segmentation algorithm. In this work, we investigate the metrics belonging to the supervised evaluation category.

Pixel-based metric is the most popular method in the supervised approach [7], [9]. It simply calculates the errors of pixels between the reference image and segmented image. And then it groups the errors into false alarm and missed detection, which are represented with precision and recall respectively. The metric considers image segmentation as a process of pixel labeling [3], thus it is not suitable for object-level evaluation. Various metrics which extend the pixel error measure have been proposed for higher-level applications such as image segmentation at object level [10], [3], video object segmentation and tracking, and image understanding [2], [11], [12].

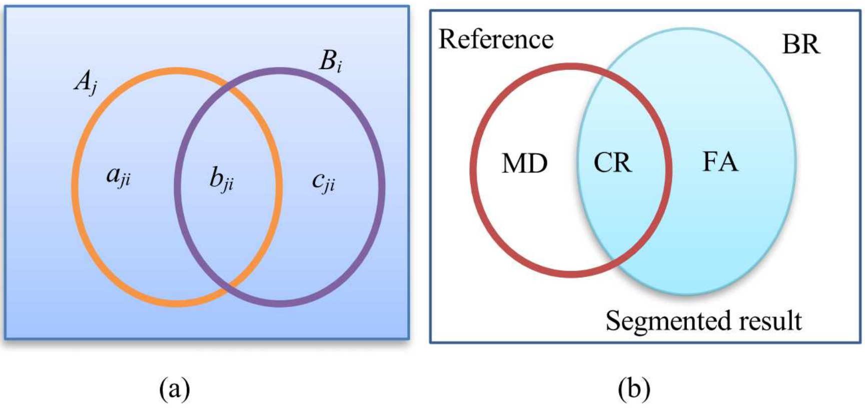
Martin et al. [10] proposed an object level error measure, including global consistency error (GCE) and local consistency error (LCE). The error metrics are very useful to quantify the consistency between segmentations manually performed by different people. However, this error measure is insensitive to over (under)-segmentation; thus it is not appropriate in segmentation applications in which the exact boundaries or sizes of the fragments are important.

To attack this problem, Polak et al. presented the object-level consistency error (OCE) [3]. The OCE quantifies the discrepancy between a segmented image and the reference image at the object level that takes into account the existence, size, position, and shape of each fragment and penalizes both over-segmentation and under-segmentation. The OCE is suitable for specific applications in which the many small objects exist in a scene and exact object size is critical in segmentation. The typical applications are the detection of crown canopies of trees, and segmentation of tar sands [3].

In this paper, we first define and describe the possible error types of image segmentation. Then we investigate the properties of the error metrics including pixel-based, Martin's, and Polak's methods by mathematical proof and experimental justification. Finally, summary and conclusion is given.

2. Error Types of Image Segmentation.

Assume $I_g = \{A_1, A_2, ..., A_M\}$ is a reference (ground-truth) image, where A_j is the jth foreground fragment (object) in I_g. Assume $I_s = \{B_1, B_2, ..., B_N\}$ is the segmented image, where B_i is the ith foreground fragment. The total numbers of pixels (called area hereafter) of the fragments A_j and B_i are denoted as $|A_j|$ and $|B_i|$, and $|A|$ and $|B|$ represent the area of I_g and I_s, respectively. According to the above definitions, we can calculate four statistics, which will be described in the following with Fig. 1 as reference. Fig. 1(b) illustrates a particular fragment pair of A_j and B_i. The b_{ji} is the area of the intersection of A_j and B_i; $a_{ji} = |A_j| - b_{ji}$, $c_{ji} = |B_i| - b_{ji}$. Note that in the following, $f(x)$ is the area of the region x; $\bar{\delta}(x) = 1 - \delta(x)$ and $\delta(x)$ is the delta function whose value equals 1 if the input is 0 and whose value equals 0 otherwise.



(a) (b)

Figure 1. Four statistics defined in this work

S1: CR (Correct Region): The area of the correctly segmented foreground region.

$$CR = f(I_g \cap I_s) = \sum_{j=1}^{M} \sum_{i=1}^{N} |A_j \cap B_i| = \sum_{j=1}^{M} \sum_{i=1}^{N} \bar{\delta}(A_j \cap B_i) b_{ji} \quad (1)$$

S2: MD (Miss Detection): The area of the miss detected foreground region.

$$MD = f(I_g \cap \overline{I_s}) = \sum_{j=1}^{M} |A_j| - CR = \sum_{j=1}^{M} |A_j| - \sum_{j=1}^{M}\sum_{i=1}^{N} \overline{\delta}(A_j \cap B_i) b_{ji} \tag{2}$$

S3: FA (False Alarm): The area of the incorrectly segmented background region.

$$FA = f(\overline{I_g} \cap I_s) = \sum_{i=1}^{N} |B_i| - CR = \sum_{i=1}^{N} |B_i| - \sum_{j=1}^{M}\sum_{i=1}^{N} \overline{\delta}(A_j \cap B_i) b_{ji} \tag{3}$$

S4: BR (Background Region): The area of correctly segmented background region.

$$BR = f(\overline{I_g} \cap \overline{I_s}) \tag{4}$$

In this paper, we classify segmentation errors into seven types, as illustrated in Fig. 2. Each type is described with the help of its corresponding graph in the following paragraphs. Note that in these graphs, the reference (ground-truth) fragments and segmented fragments are marked in gray and in yellow except the over-segmentation type (in Fig. 2(h)).

Type 1: Perfect segmentation

The segmented result is completely matched with the reference, i.e.,

$$M = N, and \begin{cases} a_{ji} = 0, b_{ji} = |A_j| = |B_i|, c_{ji} = 0, & \text{for } \overline{\delta}(A_j \cap B_i) = 1 \\ a_{ji} = 0, b_{ji} = 0, c_{ji} = 0, & \text{for } \overline{\delta}(A_j \cap B_i) = 0 \end{cases} \tag{5}$$

Type 2: Completely incorrect segmentation

The segmented result is completely mis-matched with the reference, i.e.,

$$\underset{i,j}{\forall}\ \overline{\delta}(A_j \cap B_i) = 0 \tag{6}$$

Type 3: Isolated false alarm

For simplicity without loss of generality, we define isolated false alarm as a segmented fragment which does not intersect with any reference fragment A_j exists (say B_N, marked in red in Fig. 2(d)), and the remaining reference fragments are perfectly segmented. From the definition, we have

$$M = N - 1 > 1, and \begin{cases} i \neq N, & \text{for } \overline{\delta}(A_j \cap B_i) = 1 \\ a_{jN} = |A_j|,\ b_{jN} = 0,\ c_{jN} = |B_N|, & \text{for } \overline{\delta}(A_j \cap B_i) = 0 \end{cases} \tag{7}$$

Type 4: Isolated missed detection

It contains a reference fragment that does not intersect with any segmented fragment B_i (say A_M, marked in black dash circle in Fig. 2(e)), and the remaining reference fragments are perfectly segmented. By definition, we have

$$N = M - 1 > 1, and \begin{cases} j \neq M, & \text{for } \overline{\delta}(A_j \cap B_i) = 1 \\ a_{Mi} = |A_M|,\ b_{Mi} = 0, c_{Mi} = |B_i|, & \text{for } \overline{\delta}(A_j \cap B_i) = 0 \end{cases} \tag{8}$$

Type 5: Partial false alarm /miss detection

Compared to the reference fragment, some pixels of the segmented fragment locate in the background and some pixels of a reference fragment are missed detected, as illustrated in Fig. 2(f). In this case, both partial false alarm and partial miss detection errors exist. From the definition, we have

$$\underset{i}{\exists}\ \overline{\delta}(A_j \cap B_i) = 1 \text{ and } a_{ji},\ \mathrm{b}_{ji}, \mathrm{c}_{ji} \neq 0,\ N = M - 1 > 1 \tag{9}$$

Type 6: Over segmentation

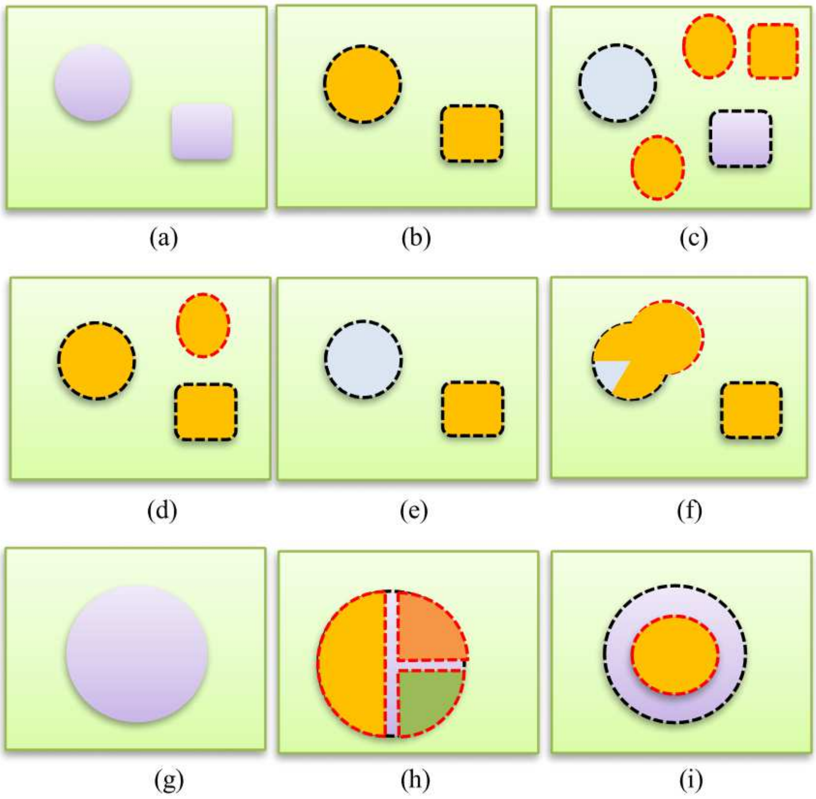


Figure 2. Error types of image segmentation, (a)Reference 1, (b)Perfect segmentation, (c)Completely incorrect segmentation, (d)Isolated false alarm, (e)Isolated missed detection, (f)Partial false alarm/miss detection, (g)Reference 2, (h)Over segmentation, (i)Under segmentation

A reference fragment A_j is segmented into more than one fragment, i.e.,

$$BS_j = \left\{ B_i \left| \bar{\delta}(A_j \cap B_i) = 1,\ i = 1, ..., k \right. \right\},\ and\ |A_j| > |BS_j| \tag{10}$$

Type 7: Under segmentation

The segmented fragment is a subset of a reference fragment, i.e., $B_i \subset A_j$ or equivalently $|B_i| < |A_j|$.

3. **Analysis of Evaluation Metrics.**

3.1. **Evaluation metrics.** Three metrics are considered in this paper. The first one is the pixel-based method. The remaining, containing Martin's metric and Polak's metric, belong to the object-based method. The pixel-based method is used as a baseline for the evaluation of other two metrics. The definition of the metrics are given in the following (refer to Fig. 1).

1. *Pixel-based metric*

   The segmentation performance is evaluated by calculating the error pixels between reference images and segmented images. Three statistics are considered here:

   (a) **PB1:**

$$\begin{aligned}\text{Precision} = \tfrac{CR}{CR+FA} &= \frac{\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}+\sum\limits_{i=1}^{N}|B_i|-\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}} \\ &= \frac{\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{\sum\limits_{i=1}^{N}|B_i|}\end{aligned} \tag{11}$$

(b) **PB2:**

$$\begin{aligned}\text{Recall} = \tfrac{CR}{CR+MD} &= \frac{\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}+\sum\limits_{j=1}^{M}|A_j|-\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}} \\ &= \frac{\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{\sum\limits_{j=1}^{M}|A_j|}\end{aligned} \tag{12}$$

(c) **PB3:**

$$\begin{aligned}F &= \tfrac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} = \tfrac{2\cdot CR}{2\cdot CR+FA+MD} \\ &= \frac{2\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{2\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}+\sum\limits_{j=1}^{M}|A_j|-\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}+\sum\limits_{i=1}^{N}|B_i|-\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}} \\ &= \frac{2\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}\end{aligned} \tag{13}$$

The precision measures the false alarm error rate; the recall measures the miss detection error rate; the F measure combines false alarm and miss detection in a single metric.

2. *Object-based metric: Martin's method*

The metric considers object (fragment) difference between the reference and segmented images. Three statistics defined in [10] are rewritten as follows (refer to Fig. 1).

(a) **MS1:**The error between fragment A_j and B_i:

$$P_{ji} = \left(1 - \frac{|A_j \cap B_i|}{|A_j|}\right) \cdot |A_j \cap B_i| = \left(1 - \frac{b_{ji}}{a_{ji} + b_{ji}}\right) \cdot b_{ji} = \frac{a_{ji}b_{ji}}{a_{ji} + b_{ji}} \tag{14}$$

(b) **MS2:**The error between fragment B_i and A_j:

$$Q_{ji} = \left(1 - \frac{|A_j \cap B_i|}{|B_i|}\right) \cdot |A_j \cap B_i| = \left(1 - \frac{b_{ji}}{b_{ji} + c_{ji}}\right) \cdot b_{ji} = \frac{b_{ji}c_{ji}}{b_{ji} + c_{ji}} \tag{15}$$

(c) **MS3:**

$$\sum_{j=1}^{M}\sum_{i=1}^{N}\min(P_{ji}, Q_{ji}) \tag{16}$$

The value of P_{ji} is determined by the missed detection error a_{ji}; the greater the missed detection error, the larger the P_{ji}. Similarly, Martin's method:

(d) **OM1:**Global Consistency Error (GCE)

$$GCE(I_g, I_s) = \tfrac{1}{n} \min \left\{ \sum_{j=1}^{M} \sum_{i=1}^{N} P_{ji}, \sum_{j=1}^{M} \sum_{i=1}^{N} Q_{ji} \right\}, \quad where\ n = \sum_{j=1}^{M} \sum_{i=1}^{N} |A_j \cap B_i| \tag{17}$$

(e) **OM2:**Local Consistency Error (LCE)

$$LCE(I_g, I_s) = \frac{1}{n} \sum_{j=1}^{M} \sum_{i=1}^{N} \min(P_{ji}, Q_{ji}) \tag{18}$$

3. *Object-based metric: Polak's method*

The OCE metric presented in [3] improves the Martin's method by exploiting weights to penalize over-segmentation and under-segmentation, which is defined in the following (refer to Fig. 1).

$$\begin{aligned} E_{g,s}(I_g, I_s) &= \sum_{j=1}^{M} \left[1 - \sum_{i=1}^{N} \tfrac{|A_j \cap B_i|}{|A_j \cup B_i|} \cdot W_{ji} \right] W_j \\ W_{ji} &= \frac{\bar{\delta}(A_j \cap B_i)|B_i|}{\sum\limits_{k=1}^{N} \bar{\delta}(A_j \cap B_k)|B_k|} \\ W_j &= \frac{|A_j|}{\sum\limits_{l=1}^{M} |A_l|} \end{aligned} \tag{19}$$

The W_j is the area ratio of the jth fragment over all fragments in the reference. The larger area ratio for a particular fragment indicates it is more important and thus the weight is larger. For a reference fragment A_j, all the segmented fragments may overlap with A_j. If the overlapping area between B_i and A_j is larger, as compared to the total overlapping area, it indicates the fragment B_i is more significant and thus W_{ji} becomes larger.

Two measures $E_{g,s}(I_g, I_s)$ and $E_{s,g}(I_s, I_g)$ in this method can be rewritten as:

(a) **PS1:**

$$E_{g,s}(I_g, I_s) = 1 - \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{b_{ji}}{a_{ji} + b_{ji} + c_{ji}} \cdot \frac{\bar{\delta}(b_{ji})(b_{ji} + c_{ji})}{\sum\limits_{k=1}^{N} \bar{\delta}(b_{jk})(b_{jk} + c_{jk})} \cdot \frac{(a_{ji} + b_{ji})}{\sum\limits_{l=1}^{M} (a_{li} + b_{li})} \tag{20}$$

(b) **PS2:**

$$E_{s,g}(I_s, I_g) = 1 - \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{b_{ji}}{a_{ji} + b_{ji} + c_{ji}} \cdot \frac{\bar{\delta}(b_{ji})(a_{ji} + b_{ji})}{\sum\limits_{l=1}^{M} \bar{\delta}(b_{li})(a_{li} + b_{li})} \cdot \frac{(b_{ji} + c_{ji})}{\sum\limits_{k=1}^{N} (b_{jk} + c_{jk})} \tag{21}$$

Polak's method combines the above two measures in a single measure OCE as

(c) **OP1:**

$$OCE(I_g, I_s) = \min\left(E_{g,s}, E_{s,g}\right) \tag{22}$$

OCE considers both $E_{g,s}(I_g, I_s)$ and $E_{s,g}(I_s, I_g)$, hence it takes both missed detection and false alarm errors into account.

3.2. **Analysis of three metrics.** The analysis is performed according to the definition of the seven types of segmentation errors in Section II. Mathematical proof and experimental justification are presented for investigating the property of the three metrics stated above. For simplicity without loss of generality, we manually generate two-level (foreground and background) ground-truth for the images downloaded from the website. Then we applied the interactive image segmentation tool developed in [13] to obtain segmentation results.

The segmentation of total 21 images are categorized into 7 error types as defined before. We then used the three error metrics to calculate the error measures for every image of each type, and the results are shown in Table I to Table VII. Each table contains several items as follows: original image, ground-truth (reference) image, segmented image, ground-truth with connected component labeling (GT_ccl), segmented with connected component labeling (Seg_ccl), relation, and error measures for the three methods. The "relation" presents the accurate segmentation and inaccurate segmentation in different colors, where yellow: correct region, blue: false alarm, green: miss detection, and white: background region.

Type 1: perfect segmentation

Deduction 1.1 (Pixel-based method): the method evaluates correctly for perfect segmentation type

Proof:

Since $b_{ji} = |A_j| = |B_i|$, we have

$$F=\frac{2\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j\cap B_i)b_{ji}}{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}=\frac{\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j\cap B_i)b_{ji}+\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M}\bar{\delta}(A_j\cap B_i)b_{ji}}{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}=\frac{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}=1$$

Deduction 1.2 (Martin's method): the method measures correctly for perfect segmentation type

Proof:

In this case, $a_{ji} = 0$, $c_{ji} = 0$, so

$P_{ji} = \frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}} = 0$, $Q_{ji} = \frac{b_{ji}\,c_{ji}}{b_{ji}+c_{ji}} = 0$,

$$\sum_{j=1}^{M}\sum_{i=1}^{N}P_{ji} = \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{a_{ji}\cdot b_{ji}}{a_{ji}+b_{ji}} = \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{0\cdot b_{ji}}{0+b_{ji}} = 0,\ \sum_{j=1}^{M}\sum_{i=1}^{N}Q_{ji} = \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{b_{ji}\cdot c_{ji}}{b_{ji}+c_{ji}} = \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{b_{ji}\cdot 0}{b_{ji}+0} = 0,$$

$$n = \sum_{j=1}^{M}\sum_{i=1}^{N}|A_j \cap B_i| = \sum_{j=1}^{M}\sum_{i=1}^{N}b_{ji},$$

and

$$GCE(I_g, I_s) = \tfrac{1}{n}\min\left\{\sum_{j=1}^{M}\sum_{i=1}^{N}P_{ji}, \sum_{j=1}^{M}\sum_{i=1}^{N}Q_{ji}\right\} = 0.$$

Since $\min(P_{ji}, Q_{ji}) = \min(0, 0) = 0$, we get

$$LCE(I_g, I_s) = \tfrac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N}\min(P_{ji}, Q_{ji})\right] = 0.$$

Thus, $GCE = LCE = 0$ for perfect segmentation.

Deduction 1.3 (Polak's method): the method measures correctly for perfect segmentation type.

Proof:

Since $a_{ji} = 0$, $c_{ji} = 0$, we have

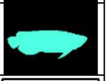
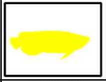
$$\begin{aligned}E_{g,s}(I_g, I_s) &= 1 - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{b_{ji}}{a_{ji}+b_{ji}+c_{ji}}\cdot\frac{\bar{\delta}(b_{ji})(b_{ji}+c_{ji})}{\sum\limits_{k=1}^{N}\bar{\delta}(b_{jk})(b_{jk}+c_{jk})}\cdot\frac{(a_{ji}+b_{ji})}{\sum\limits_{l=1}^{M}(a_{li}+b_{li})}\\ &= 1 - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{b_{ji}}{0+b_{ji}+0}\cdot\frac{1\cdot(b_{ji}+0)}{\sum\limits_{k=1}^{N}1\cdot(b_{jk}+0)}\cdot\frac{(0+b_{ji})}{\sum\limits_{l=1}^{M}(0+b_{li})}\\ &= 1 - \sum_{j=1}^{M}\sum_{i=1}^{N}1\cdot\tfrac{1}{N}\cdot\tfrac{|A_j|}{|A|} = 1 - \tfrac{1}{N}\sum_{j=1}^{M}\tfrac{N\cdot|A_j|}{|A|} = 1 - \tfrac{1}{|A|}\,|A| = 0,\end{aligned}$$

$$\begin{aligned}E_{s,g}(I_s,I_g) &= 1-\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{b_{ji}}{a_{ji}+b_{ji}+c_{ji}}\cdot\frac{\bar{\delta}(b_{ji})(a_{ji}+b_{ji})}{\sum\limits_{l=1}^{M}\bar{\delta}(b_{li})(a_{li}+b_{li})}\cdot\frac{(b_{ji}+c_{ji})}{\sum\limits_{k=1}^{N}(b_{jk}+c_{jk})}\\ &= 1-\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{b_{ji}}{0+b_{ji}+0}\cdot\frac{1\cdot(0+b_{ji})}{\sum\limits_{l=1}^{M}1\cdot(0+b_{li})}\cdot\frac{(b_{ji}+0)}{\sum\limits_{k=1}^{N}(b_{jk}+0)}\\ &= 1-\frac{1}{N\cdot|A|}\sum_{j=1}^{M}\sum_{i=1}^{N}|A_j| = 1-\frac{1}{N\cdot|A|}\sum_{j=1}^{M}N\cdot|A_j| = 1-\frac{1}{N\cdot|A|}N\cdot|A| = 0,\end{aligned}$$

$OCE(I_g,I_s) = \min\left(E_{g,s},E_{s,g}\right) = 0.$

The proofs of the above three deductions are justified by experiments, and the results are shown in Table I.

Table 1. The results for perfect segmentatoin

Original image				
Ground-truth				
Segmentation				
GT_ccl				
Seg_ccl				
Relation				
Pixel based	Precision	1.000	1.000	1.000
	Recall	1.000	1.000	1.000
	*F*_measure	1.000	1.000	1.000
[10]	*GCE*	0.000	0.000	0.000
	LCE	0.000	0.000	0.000
[3]	*OCE*	0.000	0.000	0.000

Type 2: completely inaccurate segmentation

Deduction 2.1 (Pixel-based method): the method measures correctly for completely inaccurate segmentation.

Proof:

Since $\underset{i,j}{\forall}\ \bar{\delta}(A_j\cap B_i)=0$, we have

$$F=\frac{2\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\bar{\delta}(A_j\cap B_i)b_{ji}}{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}=\frac{2\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}0\cdot b_{ji}}{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}=\frac{0}{\sum\limits_{j=1}^{M}|A_j|+\sum\limits_{i=1}^{N}|B_i|}=0$$

Deduction 2.2 (Martin's method): the measure is undefined for completely inaccurate segmentation.

Proof:

In this case, $b_{ji}=0$, so

$$P_{ji}=\frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}}=0, Q_{ji}=\frac{b_{ji}c_{ji}}{b_{ji}+c_{ji}}=0\ \sum_{j=1}^{M}\sum_{i=1}^{N}P_{ji}=\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{a_{ji}\cdot 0}{a_{ji}+0}=0,$$

$$\sum_{j=1}^{M}\sum_{i=1}^{N}Q_{ji}=\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{0\cdot c_{ji}}{0+c_{ji}}=0,$$

and $n=\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}|A_j\cap B_i|=\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}b_{ji}=0.$

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

and $LCE(I_g, I_s) = \frac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N}\min(P_{ji}, Q_{ji})\right]$ is undefined.

Deduction 2.3 (Polak's method): the method is undefined for completely inaccurate segmentation.

Proof:

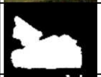
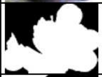
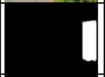
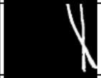
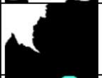
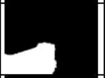
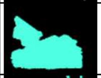
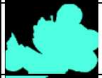
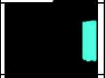
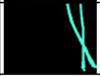
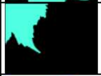
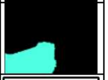
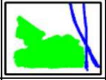
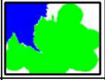
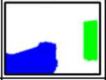
In this case, $b_{ji} = 0$, thus

$$W_{ji} = \frac{\bar{\delta}(A_j \cap B_i)|B_i|}{\sum_{k=1}^{N}\bar{\delta}(A_j \cap B_k)|B_k|} = \frac{\bar{\delta}(b_{ji})(b_{ji}+c_{ji})}{\sum_{k=1}^{N}\bar{\delta}(b_{jk})(b_{jk}+c_{jk})} = \frac{0\cdot(0+|B_i|)}{\sum_{k=1}^{N} 0\cdot(0+|B_i|)} = \frac{0}{0} \text{ is undefined.}$$

Similarly, $\sum_{l=1}^{M}\bar{\delta}(b_{li})(a_{li}+b_{li}) = 0$, thus $E_{g,s}$ and $E_{s,g}$ are undefined. Consequently, $OCE(I_g, I_s)$ is undefined.

Experiments results in Table II justify the proofs of Deductions 2.1 to 2.3.

Table 2. The results for cpmpletely inaccurate segmentatoin

Original image				
Ground-truth				
Segmentation				
GT_ccl				
Seg_ccl				
Relation				
Pixel based	Precision	0.000	0.000	0.000
	Recall	0.000	0.000	0.000
	*F*_measure	0.000	0.000	0.000
[10]	*GCE*	NaN	NaN	NaN
	LCE	NaN	NaN	NaN
[3]	*OCE*	NaN	NaN	NaN

Type 3: Isolated false alarm

Deduction 3.1 (Pixel-based method): the method evaluates correctly for isolated false alarm error.

Proof:

Since $\bar{\delta}(A_j \cap B_i) = 1 \;\; \mathrm{for} i \neq N$, $\bar{\delta}(A_j \cap B_N) = 0$, we have

$$F = \frac{2\sum_{j=1}^{M}\sum_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|} = \frac{2\sum_{j=1}^{M}\sum_{i=1}^{N-1}\bar{\delta}(A_j \cap B_i)b_{ji} + 2\sum_{j=1}^{M} 0\cdot b_{jN}}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|} = \frac{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N-1}|B_i|}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|} = \frac{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i| - |B_N|}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|}$$

$$= 1 - \frac{|B_N|}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|} = 1 - \frac{|B_N|}{2\sum_{i=1}^{N}|B_i| - |B_N|} = 1 - \frac{|B_N|}{2|B| - |B_N|}$$

The above result indicates that the F measure is related only to the area of isolated false alarm $|B_N|$, which can also be measured by precision, as justified by the experiments in Table III.

Deduction 3.2 (Martin's method): the isolated false alarm error is always missed in GCE and LCE in calculation; hence this type of errors cannot be measured by this method.

Proof:

Since $b_{jN} = 0$, we have

$$\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ji} = \sum_{j=1}^{M}\sum_{i=1}^{N-1} \frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}} + \sum_{j=1}^{M} \frac{a_{jN}b_{jN}}{a_{jN}+b_{jN}} = \sum_{j=1}^{M}\sum_{i=1}^{N-1} \frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}},$$

$$\sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji} = \sum_{j=1}^{M}\sum_{i=1}^{N-1} \frac{b_{ji}c_{ji}}{b_{ji}+c_{ji}} + \sum_{j=1}^{M} \frac{b_{jN}c_{jN}}{b_{jN}+c_{jN}} = \sum_{j=1}^{M}\sum_{i=1}^{N-1} \frac{b_{ji}c_{ji}}{b_{ji}+c_{ji}},$$

$$n = \sum_{j=1}^{M}\sum_{i=1}^{N} |A_j \cap B_i| = \sum_{j=1}^{M}\sum_{i=1}^{N} b_{ji} = \sum_{j=1}^{M}\sum_{i=1}^{N-1} b_{ji} + b_{jN} = \sum_{j=1}^{M}\sum_{i=1}^{N-1} b_{ji}$$

So,

$$\begin{aligned} GCE(I_g, I_s) &= \tfrac{1}{n}\min\left\{\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ji}, \sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji}\right\} \\ &= \tfrac{1}{n}\min\left\{\sum_{j=1}^{M}\sum_{i=1}^{N-1} P_{ji}, \sum_{j=1}^{M}\sum_{i=1}^{N-1} Q_{ji}\right\} = GCE\left(I_g, I_s - \{B_N\}\right) \end{aligned}$$

Since $\min(P_{jN}, Q_{jN}) = \min\left(\frac{a_{jN}b_{jN}}{a_{jN}+b_{jN}}, \frac{b_{jN}c_{jN}}{b_{jN}+c_{jN}}\right) = 0$, we have

$$\begin{aligned} LCE(I_g, I_s) &= \tfrac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N} \min(P_{ji}, Q_{ji})\right] = \tfrac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N-1} \min(P_{ji}, Q_{ji}) + \min(P_{jN}, Q_{jN})\right] \\ &= LCE\left(I_g, I_s - \{B_N\}\right) \end{aligned}$$

The above derivation indicates that the isolated false alarm error B_N is always missed in GCE and LCE in calculation; hence this type of errors cannot be measured by Martin's method. This is justified by the experimental results in Table III.

Deduction 3.3 (Polak's method): $OCE(I_g, I_s)$ is undefined for isolated false alarm error.

Proof:

By definition

$$E_{g,s}(I_g, I_s) = \sum_{j=1}^{M}\left[1 - \sum_{i=1}^{N} \frac{|A_j \cap B_i|}{|A_j \cup B_i|} \cdot W_{ji}\right] W_j$$

$$W_{ji} = \frac{\bar{\delta}(A_j \cap B_i)|B_i|}{\sum_{k=1}^{N} \bar{\delta}(A_j \cap B_k)|B_k|}$$

$$W_j = \frac{|A_j|}{\sum_{l=1}^{M} |A_l|}$$

We have

$$E_{g,s}(I_g, I_s) = 1 - \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{|A_j \cap B_i|}{|A_j \cup B_i|} \cdot \frac{\bar{\delta}(A_j \cap B_i)|B_i|}{\sum_{k=1}^{N} \bar{\delta}(A_j \cap B_k)|B_k|} \cdot \frac{|A_j|}{\sum_{l=1}^{M} |A_l|} = 1 - \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{|A_j \cap B_i|}{|A_j \cup B_i|} \cdot \frac{|B_i|}{|B| - |B_N|} \cdot \frac{|A_j|}{|A|}.$$

Similarly,

$$E_{s,g}(I_s, I_g) = \sum_{i=1}^{N}\left[1 - \sum_{j=1}^{M} \frac{|A_j \cap B_i|}{|A_j \cup B_i|} \cdot W_{ij}\right] W_i$$

$$W_{ij} = \frac{\bar{\delta}(A_j \cap B_i)|A_j|}{\sum_{k=1}^{M} \bar{\delta}(A_k \cap B_i)|A_k|}$$

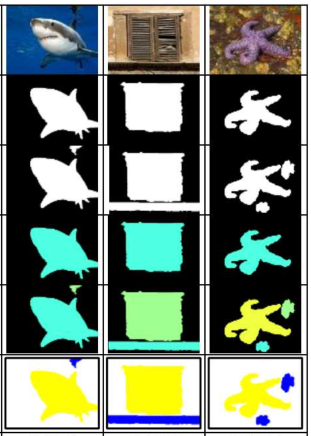
$$W_i = \frac{|B_i|}{\sum_{l=1}^{N} |B_l|}$$

Assume B_N is a fragment of isolated false alarm error, then $\bar{\delta}(A_j \cap B_N) = 0$ for all j.

We have $W_{Nj} = \frac{\bar{\delta}(A_j \cap B_N)|A_j|}{\sum_{k=1}^{M} \bar{\delta}(A_k \cap B_N)|A_k|} = \frac{0}{0}$ is undefined.

Thus $E_{s,g}(I_s, I_g)$ is undefined, and $OCE(I_g, I_s) = min(E_{g,s}, E_{s,g})$ is undefined accordingly, which is justified in Table III.

Table 3. The results for isolated false alarm



Pixel based	Precision	0.972	0.780	0.827
	Recall	1.000	1.000	1.000
	*F*_measure	0.986	0.876	0.905
[10]	*GCE*	0.000	0.000	0.000
	LCE	0.000	0.000	0.000
[3]	*OCE*	NaN	NaN	NaN

Type 4: Isolated miss detection

Deduction 4.1(Pixel-based method): pixel-based method measures correctly for isolated missed detection error.

Proof:

Since $\bar{\delta}(A_j \cap B_i) = 1$ for $j \neq M$, $\bar{\delta}(A_M \cap B_i) = 0$, we have

$$F = \frac{2\sum_{j=1}^{M-1}\sum_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji} + 2\sum_{i=1}^{N} 0 \cdot b_{Mi}}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|} = \frac{\sum_{i=1}^{N}|B_i| + \sum_{j=1}^{M-1}|A_j|}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|} = \frac{\sum_{i=1}^{N}|B_i| + \sum_{j=1}^{M}|A_j| - |A_M|}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|}$$

$$= 1 - \frac{|A_M|}{2\sum_{j=1}^{M}|A_j| - |A_M|} = 1 - \frac{|A_M|}{2|A| - |A_M|}$$

The experimental result in Table IV indicates the F measure evaluates correctly for isolated missed detection error and is determined only by recall, without any relation to precision.

Deduction 4.2 (Martin's method): the isolated miss detection error is always lost in GCE and LCE in calculation; hence the method is unable to measure this type of errors.

Proof:

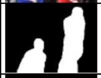
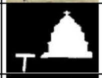
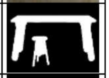
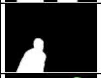
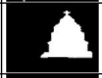
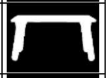
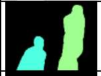
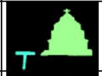
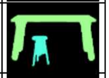
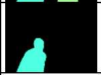
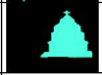
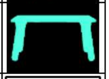
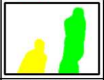
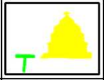
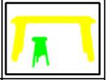
The proof is analogous to that of Deduction 3.2 just by switching a_{ji} and c_{ji}. We obtain the result that the isolated miss detection error A_M is always lost in the calculation of GCE and LCE; thus the Martin's method is unable to evaluate the error. Experimental results are shown in Table IV.

Deduction 4.3 (Polak's method): $OCE(I_g, I_s)$ is undefined for isolated missed detection error.

Proof:

By definition

Table 4. The results for isolated miss detection

Original image				
Ground-truth				
Segmentation				
GT_ccl				
Seg_ccl				
Relation				
Pixel based	Precision	1.000	1.000	1.000
	Recall	0.343	0.924	0.809
	*F*_measure	0.511	0.961	0.894
[10]	*GCE*	0.000	0.000	0.000
	LCE	0.000	0.000	0.000
[3]	*OCE*	NaN	NaN	NaN

$$E_{g,s}(I_g, I_s) = \sum_{j=1}^{M}\left[1 - \sum_{i=1}^{N}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot W_{ji}\right]W_j$$

$$W_{ji} = \frac{\bar{\delta}(A_j \cap B_i)|B_i|}{\sum_{k=1}^{N}\bar{\delta}(A_j \cap B_k)|B_k|}$$

$$W_j = \frac{|A_j|}{\sum_{l=1}^{M}|A_l|}$$

Assume A_M is a fragment of isolated miss detection, then $\bar{\delta}(A_M \cap B_i) = 0$ for all i. In calculation of $E_{g,s}(I_g, I_s)$, we have $W_{Mi} = \frac{\bar{\delta}(A_M \cap B_i)|B_i|}{\sum_{k=1}^{N}\bar{\delta}(A_M \cap B_k)|B_k|} = \frac{0}{0}$ is undefined.

Consequently, $E_{g,s}(I_g, I_s)$ and $OCE(I_g, I_s)$ are undefined, which is justified by Table IV.

Type 5: Partial false alarm/miss detection

Deduction 5.1 (Pixel-based method): it can measure both partial false alarm and miss detection errors correctly, and the F value is proportional to the error value.

Proof:

$$F = \frac{2\sum_{j=1}^{M}\sum_{i=1}^{N}\bar{\delta}(A_j \cap B_i)b_{ji}}{\sum_{j=1}^{M}|A_j| + \sum_{i=1}^{N}|B_i|} \in [0, 1]$$

Deduction 5.2 (Martin's method): $GCE(LCE)$ always outputs the smaller of the two errors when miss detection (a_{ji}) and false alarm (c_{ji}) errors exist simultaneously. Thus the method will generate wrong measure.

Proof:

From the definition, P_{ji} is determined by missed detection (a_{ji}). The larger a_{ji}, the greater the P_{ji}. On the contrary, the false alarm (c_{ji}) determines the value of Q_{ji}, and larger c_{ji} yields larger Q_{ji}. We prove the deduction with two extreme cases as follows.

(a) Serious false alarm $(c_{ji} >> a_{ji})$

$$\sum_{j=1}^{M}\sum_{i=1}^{N}P_{ji} = \sum_{j=1}^{M}\sum_{i=1}^{N}b_{ji} - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{(b_{ji})^2}{a_{ji}+b_{ji}} < \sum_{j=1}^{M}\sum_{i=1}^{N}b_{ji} - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{(b_{ji})^2}{b_{ji}+c_{ji}} = \sum_{j=1}^{M}\sum_{i=1}^{N}Q_{ji}.$$

Thus $GCE(I_g, I_s) = \frac{1}{n}\min\left\{\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ji}, \sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji}\right\} = \frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ji} = \frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{N} \frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}}$.

Since $P_{ji} = \frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}} = \frac{1}{\frac{1}{a_{ji}}+\frac{1}{b_{ji}}} < \frac{1}{\frac{1}{c_{ji}}+\frac{1}{b_{ji}}} = \frac{c_{ji}b_{ji}}{c_{ji}+b_{ji}} = Q_{ji}$,

we have $LCE(I_g, I_s) = \frac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N}\min(P_{ji}, Q_{ji})\right] = \frac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N-1} P_{ji}\right] = \frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}}$.

The output is determined by a_{ji} instead of c_{ji}.

(b) Serious miss detection ($a_{ji} >> c_{ji}$)

$\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ji} = \sum_{j=1}^{M}\sum_{i=1}^{N} b_{ji} - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{(b_{ji})^2}{a_{ji}+b_{ji}} > \sum_{j=1}^{M}\sum_{i=1}^{N} b_{ji} - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{(b_{ji})^2}{b_{ji}+c_{ji}} = \sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji}$.

So $GCE(I_g, I_s) = \frac{1}{n}\min\left\{\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ji}, \sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji}\right\} = \frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji} = \frac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{c_{ji}b_{ji}}{c_{ji}+b_{ji}}\right]$.

Since $P_{ji} = \frac{a_{ji}b_{ji}}{a_{ji}+b_{ji}} = \frac{1}{\frac{1}{a_{ji}}+\frac{1}{b_{ji}}} > \frac{1}{\frac{1}{c_{ji}}+\frac{1}{b_{ji}}} = \frac{c_{ji}b_{ji}}{c_{ji}+b_{ji}} = Q_{ji}$,

we have $LCE(I_g, I_s) = \frac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N}\min(P_{ji}, Q_{ji})\right] = \frac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji}\right] = \frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{c_{ji}b_{ji}}{c_{ji}+b_{ji}}$.

The output is determined by c_{ji} instead of a_{ji}.

The above derivation indicates that when both miss detection error and false alarm error exist, we have

(a) $LCE(I_g, I_s) = GCE(I_g, I_s)$, and

(b) Either GCE or LCE outputs the smaller of these two errors.

The dominant source of error is always missed. Therefore, the measure is obviously wrong. The results are justified by Table V.

Deduction 5.3 (Polak's method): $OCE(I_g, I_s)$ can measure errors correctly when miss detection and false alarm errors exist simultaneously. However, it will have $OCE(I_g, I_s) = E_{g,s} = E_{s,g}$, which means OCE cannot distinguish partial miss detection error from partial false alarm error.

Proof:

(a) Serious false alarm ($c_{ji} >> a_{ji}$)

Since $\exists\bar{\delta}(A_j \cap B_i) = 1$, we have

$$E_{g,s}(I_g, I_s) = \sum_{j=1}^{M}\left(\frac{|A_j|}{\sum_{l=1}^{M}|A_l|}\right) - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot\frac{\bar{\delta}(A_j \cap B_i)|B_i|}{\sum_{k=1}^{N}\bar{\delta}(A_j \cap B_k)|B_k|}\cdot\frac{|A_j|}{\sum_{l=1}^{M}|A_l|} = 1 - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot\frac{|B_i|}{|B|}\cdot\frac{|A_j|}{|A|},$$

$$E_{s,g}(I_s, I_g) = \sum_{i=1}^{N}\left(\frac{|B_i|}{\sum_{l=1}^{N}|B_l|}\right) - \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot\frac{\bar{\delta}(A_j \cap B_i)|A_j|}{\sum_{k=1}^{M}\bar{\delta}(A_k \cap B_i)|A_k|}\cdot\frac{|B_i|}{\sum_{l=1}^{N}|B_l|} = 1 - \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot\frac{|A_j|}{|A|}\cdot\frac{|B_i|}{|B|}.$$

Hence, $E_{g,s}(I_g, I_s) = E_{s,g}(I_s, I_g)$, and $OCE(I_g, I_s) = \min(E_{g,s}, E_{s,g}) = E_{g,s} = E_{s,g}$.

For the special case, $M = N = 1$, we have

$$\begin{aligned} OCE(I_g, I_s) &= E_{g,s} = E_{s,g} = 1 - \frac{|A_j \cap B_i|}{|A_j \cup B_i|} = 1 - \frac{b_{11}}{a_{11}+b_{11}+c_{11}} = \frac{a_{11}+c_{11}}{a_{11}+b_{11}+c_{11}} = \frac{\frac{a_{11}}{c_{11}}+1}{\frac{a_{11}}{c_{11}}+\frac{b_{11}}{c_{11}}+1} \\ &= \frac{1}{\frac{b_{11}}{c_{11}}+1} = \frac{c_{11}}{b_{11}+c_{11}}(c_{11} >> a_{11}) \end{aligned}$$

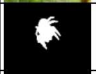
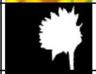
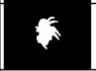
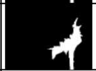
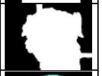
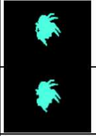
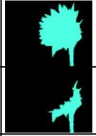
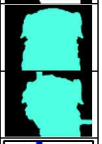
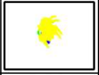
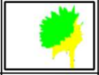
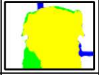
The result implies that OCE is determined by false alarm error c_{ji}, which is correct.

(b) Serious miss detection ($a_{ji} >> c_{ji}$)

Since $\exists\bar{\delta}(A_j \cap B_i) = 1$, we have

$$E_{g,s}(I_g, I_s) = 1 - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot\frac{\bar{\delta}(A_j \cap B_i)|B_i|}{\sum_{k=1}^{N}\bar{\delta}(A_j \cap B_k)|B_k|}\cdot\frac{|A_j|}{\sum_{l=1}^{M}|A_l|} = 1 - \sum_{j=1}^{M}\sum_{i=1}^{N}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot\frac{|B_i|}{|B|}\cdot\frac{|A_j|}{|A|},$$

Table 5. The results for partial false alarm/miss detection

Original image				
Ground-truth				
Segmentation				
GT_ccl				
Seg_ccl				
Relation				
Pixel based	Precision	0.980	0.998	0.961
	Recall	0.971	0.332	0.889
	*F*_measure	0.975	0.499	0.923
[10]	*GCE*	0.021	0.002	0.039
	LCE	0.021	0.002	0.039
[3]	*OCE*	0.048	0.668	0.141

$$E_{s,g}(I_s,I_g)=\sum_{i=1}^{N}\left(\frac{|B_i|}{\sum\limits_{l=1}^{N}|B_l|}\right)-\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|A_j\cap B_i|}{|A_j\cup B_i|}\cdot\frac{\bar{\delta}(A_j\cap B_i)|A_j|}{\sum\limits_{k=1}^{M}\bar{\delta}(A_k\cap B_i)|A_k|}\cdot\frac{|B_i|}{\sum\limits_{l=1}^{N}|B_l|}=1-\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|A_j\cap B_i|}{|A_j\cup B_i|}\cdot\frac{|A_j|}{|A|}\cdot\frac{|B_i|}{|B|}.$$

So $E_{g,s}(I_g,I_s)=E_{s,g}(I_s,I_g)$, and $OCE(I_g,I_s)=\min\left(E_{g,s},E_{s,g}\right)=E_{s,g}=E_{g,s}$

For the special case, $M=N=1$, we have

$$E_{g,s}=E_{s,g}=1-\frac{|A_j\cap B_i|}{|A_j\cup B_i|}=1-\frac{b_{11}}{a_{11}+b_{11}+c_{11}}=\frac{a_{11}+c_{11}}{a_{11}+b_{11}+c_{11}}=\frac{1+\frac{c_{11}}{a_{11}}}{1+\frac{b_{11}}{a_{11}}+\frac{c_{11}}{a_{11}}}$$
$$\simeq\frac{1}{1+\frac{b_{11}}{a_{11}}}=\frac{a_{11}}{a_{11}+b_{11}}(a_{11}>>c_{11}).$$

This also implies that OCE is determined by miss detection error a_{ji}, which is correct.

From the derivation above, we found $E_{s,g}$ is always equal to $E_{g,s}$. This means the method cannot distinguish partial miss detection error from partial false alarm error.

Type 6: Over segmentation

Deduction 6.1 (Pixel-based method): the method evaluates correctly for the over segmentation and the F measure is always greater than Recall.

Proof:

Assume the fragment A_1 in I_s is partitioned into k fragments which composite a set BS, $BS=\left\{B_i\left|\bar{\delta}(A_1\cap B_i)=1,\ i=1,...,k\right.\right\}$, and $|A_1|>|BS|$, $a_{1i},\ b_{1i}\neq 0$, $c_{1i}=0$. We have $F=\frac{2\sum\limits_{i=1}^{k}\bar{\delta}(A_1\cap B_i)b_{1i}}{|A_1|+|BS|}=\frac{2\sum\limits_{i=1}^{k}b_{1i}}{|A_1|+|BS|}=\frac{2|BS|}{|A_1|+|BS|}>\frac{|BS|}{|A_1|}=\frac{CR}{CR+MD}=$ Recall.

Over segmentation is a kind of miss detection error with no false alarm. Thus the precision is equal to one, and the F value is determined only by the recall, as shown in Table VI.

Deduction 6.2 (Martin's method): GCE and LCE are always equal to zero and not able to measure over-segmentation error.

Proof:

Using the same assumption above, we have

$$\sum_{j=1}^{1}\sum_{i=1}^{k}P_{ji}=\sum_{i=1}^{k}\left[\left(1-\frac{|A_1\cap B_i|}{|A_1|}\right)\cdot|A_1\cap B_i|\right]=\sum_{i=1}^{k}\left(1-\frac{b_{1i}}{a_{1i}+b_{1i}}\right)\cdot b_{1i}=\sum_{i=1}^{k}b_{1i}+\sum_{i=1}^{k}\frac{(b_{1i})^2}{a_{1i}+b_{1i}}$$
$$=|BS|+\sum_{i=1}^{k}\frac{(b_{1i})^2}{|A_1|},$$

$$\sum_{j=1}^{1}\sum_{i=1}^{k} Q_{ji} = \sum_{i=1}^{k}\left[\left(1-\frac{|A_1\cap B_i|}{|B_j|}\right)\cdot|A_1\cap B_i|\right] = \sum_{i=1}^{k}\left(1-\frac{b_{1i}}{b_{1i}+c_{1i}}\right)\cdot b_{1i} = 0,$$

and $\min(P_{1i}, Q_{1i}) = \min\left(\frac{a_{1i}b_{1i}}{a_{1i}+b_{1i}}, \frac{b_{1i}c_{1i}}{b_{1i}+c_{1i}}\right) = \min\left(\frac{a_{1i}b_{1i}}{a_{1i}+b_{1i}}, 0\right) = 0$

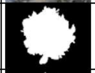
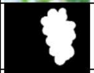
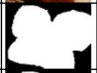
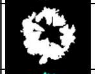
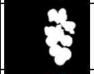
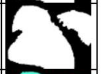
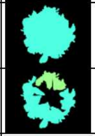
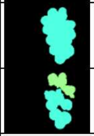
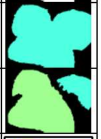
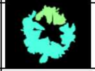
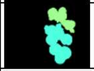
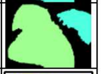
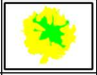
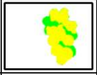
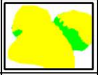
The reference fragment is fixed, i.e., $|A_1|$ is constant, thus

$$GCE(I_g, I_s) = \frac{1}{n}\min\left\{\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ji}, \sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ji}\right\} = 0,$$

$$LCE(I_g, I_s) = \frac{1}{n}\left[\sum_{j=1}^{M}\sum_{i=1}^{N}\min(P_{ji}, Q_{ji})\right] = 0$$

We conclude that in over-segmentation case $GCE = LCE = 0$, which means Martin's method is not able to measure over segmentation error. The experimental results shown in Table VI justify this.

Table 6. The results for over segmentation

Original image				
Ground-truth				
Segmentation				
GT_ccl				
Seg_ccl				
Relation				
Pixel based	Precision	1.000	1.000	1.000
	Recall	0.771	0.796	0.886
	*F*_measure	0.871	0.887	0.940
[10]	*GCE*	0.000	0.000	0.000
	LCE	0.000	0.000	0.000
[3]	*OCE*	0.488	0.554	0.382

Deduction 6.3 (Polak's method): OCE can measure over-segmentation error correctly

Proof:

Using the same assumption in Deduction 6.1, we have

$$E_{g,s}(I_g, I_s) = 1-\sum_{j=1}^{1}\sum_{i=1}^{k}\frac{|A_1\cap B_i|}{|A_1\cup B_i|}\cdot\frac{\bar{\delta}(A_1\cap B_i)|B_i|}{\sum\limits_{l=1}^{k}\bar{\delta}(A_1\cap B_l)|B_l|}\cdot\frac{|A_1|}{\sum\limits_{m=1}^{1}|A_m|} = 1-\sum_{i=1}^{k}\frac{|B_i|}{|A_1|}\cdot\frac{|B_i|}{|BS|} = 1-\frac{\sum\limits_{i=1}^{k}(|B_i|)^2}{|A_1|\cdot|BS|},$$

$$E_{s,g}(I_s, I_g) = \sum_{i=1}^{k}\left(\frac{|B_i|}{\sum\limits_{l=1}^{k}|B_l|}\right) - \sum_{i=1}^{k}\sum_{j=1}^{1}\frac{|A_j\cap B_i|}{|A_j\cup B_i|}\cdot\frac{\bar{\delta}(A_j\cap B_i)|A_j|}{\sum\limits_{k=1}^{M}\bar{\delta}(A_k\cap B_i)|A_k|}\cdot\frac{|B_i|}{\sum\limits_{l=1}^{N}|B_l|} = 1-\sum_{i=1}^{k}\frac{|A_1\cap B_i|}{|A_1\cup B_i|}\cdot\frac{|B_i|}{|BS|}$$

$$= 1-\sum_{i=1}^{k}\frac{|B_i|}{|A_1|}\cdot\frac{|B_i|}{|BS|} = 1-\frac{\sum\limits_{i=1}^{k}(|B_i|)^2}{|A_1|\cdot|BS|}$$

So $OCE(I_g, I_s) = E_{g,s}(I_g, I_s) = E_{s,g}(I_s, I_g)$.

The result indicates that OCE is proportional to the amount of over-segmentation error.

Type 7: Under segmentation

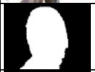
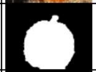
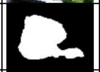
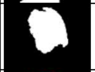
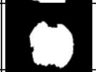
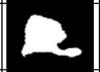
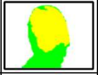
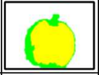
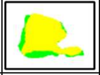
Deduction 7.1 (Pixel-based method): the method evaluates under-segmentation correctly and the F measure is greater than *recall*.

Proof:

Assume the fragment A_1 in I_s is segmented into a smaller fragment B_1, i.e., $B_1 \subset A_1$ and $|A_1| > |B_1|$. We have $F = \frac{2\bar{\delta}(A_1 \cap B_1)b_{11}}{|A_1|+|B_1|} = \frac{2|B_1|}{|A_1|+|B_1|} > \frac{|B_1|}{|A_1|} = \frac{CR}{CR+MD} = \text{Recall}$.

Under-segmentation is a kind of miss detection error. The larger the miss detection ($|A_1| > |B_1|$), the smaller the F measure, as shown in Table VII.

TABLE 7. The results for under segmentation

Original image				
Ground-truth				
Segmentation				
GT_ccl				
Seg_ccl				
Relation				
Pixel based	Precision	1.000	1.000	1.000
	Recall	0.590	0.746	0.803
	*F*_measure	0.742	0.854	0.891
[10]	*GCE*	0.000	0.000	0.000
	LCE	0.000	0.000	0.000
[3]	*OCE*	0.407	0.252	0.197

Deduction 7.2 (Martin's method): GCE and LCE are always equal to zero and not able to measure under-segmentation error.

Proof:

Using the same assumption as above, we have

$$\sum_{j=1}^{1}\sum_{i=1}^{1} P_{ji} = P_{11} = \left(1 - \frac{|A_1 \cap B_1|}{|A_1|}\right) \cdot |A_1 \cap B_1| = \left(1 - \frac{b_{11}}{a_{11}+b_{11}}\right) \cdot b_{11} = |BS| + \frac{(b_{1i})^2}{|A_1|},$$

$$\sum_{j=1}^{1}\sum_{i=1}^{1} Q_{ji} = Q_{11} = \left(1 - \frac{|A_1 \cap B_1|}{|B_1|}\right) \cdot |A_1 \cap B_1| = \left(1 - \frac{b_{11}}{b_{1i}+c_{11}}\right) \cdot b_{11} = 0,$$

and $\min(P_{11}, Q_{11}) = \min\left(\frac{a_{11}b_{11}}{a_{11}+b_{11}}, \frac{b_{11}c_{11}}{b_{11}+c_{11}}\right) = \min\left(\frac{a_{11}b_{11}}{a_{11}+b_{11}}, 0\right) = 0$.

The reference fragment is fixed, i.e., $|A_1|$ is constant, thus

$$GCE(I_g, I_s) = \frac{1}{n}\min\left\{\sum_{j=1}^{1}\sum_{i=1}^{1} P_{ji}, \sum_{j=1}^{1}\sum_{i=1}^{1} Q_{ji}\right\} = \frac{1}{n}\min\{P_{11}, Q_{11}\} = 0,$$

$$LCE(I_g, I_s) = \frac{1}{n}\left[\min(P_{11}, Q_{11})\right] = 0.$$

We conclude that in under-segmentation $GCE = LCE = 0$, which means Martin's method is not able to measure under-segmentation error, which is also justified by experiments shown in Table VII.

Deduction 7.3 (Polak's method): OCE measures under-segmentation error correctly.

Proof:

Assume the fragment A_1 in I_s is segmented into a smaller fragment B_1, i.e., $B_1 \subset A_1$ and $|A_1| > |B_1|$. In this case, we have $a_{11},\ b_{11} \neq 0$, $c_{11} = 0$. Thus

$$E_{g,s}(I_g, I_s) = 1 - \sum_{j=1}^{1}\sum_{i=1}^{1} \frac{|A_1 \cap B_i|}{|A_1 \cup B_i|} \cdot \frac{\bar{\delta}(A_1 \cap B_i)|B_i|}{\sum_{l=1}^{1}\bar{\delta}(A_1 \cap B_l)|B_l|} \cdot \frac{|A_1|}{\sum_{m=1}^{1}|A_m|} = 1 - \frac{|B_1|}{|A_1|} \cdot \frac{|B_1|}{|B_1|} = 1 - \frac{|B_1|}{|A_1|},$$

Table 8. Comparison of error measure abilities for three error metrics

Error types	**Pixel based**	**Martin's metric**	**Polak's metric**
Perfect segmentation	√	√	√
Completely inaccurate segmentation	√	×	×
Isolated false alarm	√	×	×
Isolated missed detection	√	×	×
Partial false alarm/ missed detection	√	×	□
Over-segmentation	√	×	√
Under-segmentation	√	×	√

$$\begin{aligned} E_{s,g}(I_s, I_g) &= \sum_{i=1}^{1}\left(\frac{|B_i|}{\sum\limits_{l=1}^{1}|B_l|}\right) - \sum_{i=1}^{1}\sum_{j=1}^{1}\frac{|A_j \cap B_i|}{|A_j \cup B_i|}\cdot\frac{\bar{\delta}(A_j \cap B_i)|A_j|}{\sum\limits_{k=1}^{1}\bar{\delta}(A_k \cap B_i)|A_k|}\cdot\frac{|B_i|}{\sum\limits_{l=1}^{1}|B_l|} = 1 - \frac{|A_1 \cap B_1|}{|A_1 \cup B_1|}\cdot\frac{|B_1|}{|B_1|} \\ &= 1 - \frac{|B_1|}{|A_1|} \end{aligned}.$$

So we have $OCE(I_g, I_s) = E_{g,s}(I_g, I_s) = E_{s,g}(I_s, I_g)$.

The result indicates that OCE is proportional to the amount of under-segmentation error.

4. **Summary and Conclusion.** In this paper, we have investigated the properties of three evaluation methods for image segmentation quality: pixel-based method, Martions method and Polaks method through mathematical proof and experimental justification. The results are summarized in Table VIII, where the "√" and "×" denotes the method can and cannot measure the error types, respectively. The special marker "□" represents the method is able to measure the error type, but unable to distinguish partial false alarm error from partial missed detection error. This paper shows that the object-based metrics have several deficiencies although they are suitable for object-level evaluation. In addition, Polak's method is better than Martin's method, as expected.

REFERENCES

[1] S. Shantaiya, K. Verma, and K. Mehta, A survey on approaches of object detection, *International Journal of Computer Applications*, vol. 65, no. 18, pp.14-20, 2013.

[2] C. Kim, and J. N. Hwang, Fast and automatic video object segmentation and tracking for content-based applications, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp.122-129, 2002.

[3] M. Polak, H. Zhang, and M. H. Pi, An evaluation metric for image segmentation of multiple objects, *Image and Vision Computing*, vol. 27, pp. 1223-1227, 2009.

[4] C. E. Erdem, A. M. Tekalp, and B. Sankur, Metrics for performance evaluation of video object segmentation and tracking without ground-truth, *Proc. of IEEE International Conference on Image Processing*, vol. 2, pp. 69-72, 2010.

[5] Y. J. Zhang, A survey on evaluation methods for image segmentation, *Pattern Recognition*, vol. 29, pp. 1335-1346, 1996.

[6] Y. J. Zhang, A survey on evaluation methods for image segmentation, *Proc. of the International Symposium on Signal processing and its applications*, pp. 13-16, 2001.

[7] H. Zhang, J. E. Fritts, and S. A. Goldman, Image segmentation evaluation: A survey of unsupervised methods, *Computer Vision and Image Understanding*, vol. 110, pp. 260-280, 2008.

[8] H. Li, J. Cai, T. N. A. Nguyen, and J. Zheng, A benchmark for semantic image segmentation, *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp.1-6, 2013.

[9] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, Contour detection and hierarchical image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 898-916, 2011.

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik, A database of human segmented natural images and its application to evaluating algorithms and measuring ecological statistics, *Proc. of the International Conference on Computer Vision*, pp. 416V423, 2001.

[11] B. Hemery, H. Laurent, and C. Rosenberger, Evaluation metric for image understanding, *Proc. of the International Conference on Image Processing*, pp. 4381-4384, 2009.

[12] P. L. Correia and F. Pereira, Objective evaluation of video segmentation quality, *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 186-200, 2003.

[13] T. N. A. Nguyen, J. Cai, J. Zhang, and J. Zheng, Robust interactive image segmentation using convex active contours, *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3734-3743, 2012.