2015-08

An adjustable error measure for image segmentation evaluation

Conference on Graphics, Patterns and Images, 28th, 2015, Salvador.
http://www.producao.usp.br/handle/BDPI/49194

An Adjustable Error Measure for Image Segmentation Evaluation

Oscar Cuadros Linares*, Glenda Botelho†, Francisco Rodrigues* and João Batista Neto*
*Instituto de Ciências Matemáticas e de Computação (ICMC)
University of São Paulo (USP)
São Carlos, Brazil
Email: {ocuadros, francisco, jbatista}@icmc.usp.br
†Universidade Federal do Tocantins (UFT)
Palmas-TO, Brazil
Email: glendabotelho@uft.edu.br

***Abstract*—Due to the subjective nature of the segmentation process, quantitative evaluation of image segmentation methods is still a difficult task. Humans perceive image objects in different ways. Consequently, human segmentations may come in different levels of refinement, ie, under- and over-segmentations. Popular segmentation error measures in the literature (Arbelaez and OCE) are supervised methods (also called empirical discrepancy methods), in which error is computed by comparing objects in segmentations with a reference (ground-truth) image produced by humans. Since reference images can be many, the key issue for a segmentation error measure is to be consistent in the presence of both under- and over-segmentation. In general, the term consistency refers to the ability of the error measure to be low, when comparing similar segmentations, or high, when faced with different segmentations, while capturing under- or over-segmentations. In this paper we propose a new object-based empirical discrepancy error measure, called Adjustable Object-based Measure (AOM). We introduce a penalty parameter which gives the method the ability to be more (or less) responsive in the presence of over-segmentation. Hence, we extend the notion of consistency so as to include the application's need in the process. Some applications require segmentation to be extremely accurate, hence under- or over-segmentation should be well penalised. Others, do not. By changing the penalty parameter, AOM can deliver more consistent results not only in reference to the under- or over-segmentation issue alone, but also according to the nature of the application. We compare our method with Arbelaez (used as standard measure in the benchmark of Berkeley Segmentation Image Dataset) and OCE. Our results show that AOM not only is more consistent in the presence of over-segmentation, but is also faster to compute. Unlike Arbelaez and OCE, AOM also satisfies the metric axiom of symmetry.**

***Keywords*-error measure; metric; evaluation; image segmentation;**

I. Introduction

The importance of image segmentation has long been acknowledged for many image, video and computer vision applications. In spite of the great effort towards the development of segmentation algorithms, which resulted in a huge number of efficient methods, less attention has been paid to answer the following questions: what is a correct segmentation? How can we quantitatively evaluate segmentation (and segmentation methods) given its subjective nature?

Evaluation methods for image segmentation can be categorized as either **subjective** or **objective** [1]. Subjective evaluation, conducted by human observers, generally depends on their own standards for assessing the quality of segmentation and may yield distinct results according to the order in which segmentation is observed, which introduces an undesirable bias factor to the evaluation process. Objective evaluation, in its turn, can be divided into **system-level** and **direct** evaluation methods. System-based methods are more common in systems or applications that employ different segmentation methods. The best segmentation method is the one that produces the best empirical results for the application. Shin et al [2], for example, evaluate edge-based methods in an object recognition system. The major drawback of such approach is that is simply tells that a segmentation method is more desirable for a certain application, ie, it is an indirect evaluator.

On the other hand, direct evaluation methods analyse segmentation methods independently. They can be further divided into **Analytical** and **Empirical** evaluation methods. Analytical methods [3], [4] take certain properties from the segmentation algorithms, such as complexity, efficiency and/or processing strategy (parallelism, sequentiality and iterations) to assess segmentation quality. As such, this approach can only evaluate the algorithmic and implementation properties of the segmentation methods and are not effective at telling the difference in performance of segmentation algorithms.

Empirical evaluation methods, in turn, can be categorized into **supervised** or **unsupervised** methods, based on whether the method requires a ground-truth (or gold standard) reference image or not. The former are also referred to as **empirical discrepancy** methods, while the latter are referred to as **empirical goodness** methods [3]. Empirical goodness or unsupervised methods have do not require the costly task of producing a ground-truth image (normally a manual procedure performed by a human expert) for comparison with the segmentation, and employs the original image and the segmented data. "Goodness" can be quantitatively expressed as a measure of uniformity within regions, contrast between regions or even shape of segmented regions. However, without a reference image, goodness may not be objective [5]. Moreover, for

meaningful results, an appropriate model of "goodness" would be required and should be embedded in the algorithm itself.

Conversely, empirical discrepancy or supervised methods explicitly compute the error between the segmented image and a reference (ground-truth) image. The greatest benefit of empirical discrepancy over its counterparts is that the error computed between a segmented image and a reference ground-truth provides a finer resolution of evaluation, that is, the magnitude of the error represents the level of agreement and/or disagreement. Earlier discrepancy methods conducted evaluation based on the number of mis-segmented pixels [6], [7], position of mis-segmented pixels [8] and then extended to consider the number of objects in the image [9], [10].

Despite the potential benefits of empirical discrepancy methods, we cannot ignore certain issues while comparing them against reference images. For most images, especially natural scene ones, we cannot guarantee a unique manually-segmented reference image. On the contrary, segmentations of a single image produced by different people are likely to be different as humans perceive scenes differently or perform segmentation in different granularities. Therefore, comparisons based on such references are somehow subjective.

For that reason, the three most import empirical discrepancy evaluation measures, namely Martin[11], OCE[5] and Arbelaez[12] introduce mechanisms to curb or penalise under- and over-segmentation. They are all object-based discrepancy measures and seek to quantify the consistency between segmentations manually performed by different people of the same image, who perceive the scene at different granularities or scales.

However, they all exhibit some drawbacks. The Martin measure [11], for instance, is not appropriate for applications in which details of real object boundaries are important, as it does not penalise over-segmentations [5]. As a result, this measure will consider segmented images with different degrees of granularities as being similar. The OCE error measure [5], which is an extension of Martin measure, has added a penalty factor to over-segmentation. First, it does not satisfy the symmetry axiom. As a result, OCE requires a double computation of the error for two segmented images I_s and I_g. The minimum value is then picked as the correct measure ($OCE(I_s, I_g) = \min(E_{g,s}, E_{s,g})$). This increases computation times substantially. Moreover, it is very sensitive to slight variations in segmentations, over penalising segmented images which look very similar. Arbelaez measure also does not satisfy the symmetry axiom and no solution for that is presented. Moreover, it does not penalise over-segmentation.

Contributions: We propose a new object-based empirical discrepancy error measure, called Adjustable Object-based Measure (AOM), which computes discrepancies (or similarities) between a reference image and a segmented image at the object level. Contributions include:

- Ability to adjust the penalty degree to compensate for the inherent subjectiveness of supervised evaluation methods. We introduce a parameter α which allows our method to become tolerant to refinement in a flexible way.
- The error measure satisfies the metric axioms of identity and symmetry. Exhaustive experimentation supports the conjecture that it also satisfies the triangle inequality.
- It is 2.716 times faster to compute than Abelaez, which is currently the most employed empirical discrepancy method, and 104.781 times faster than OCE.

This paper is organised as follows. Section II describes the error measures proposed by Martin, Polak and Arbelaez. Section III describes the proposed AOM error measure. Experimental results are given in section IV. Conclusions and future work are finally described in section V.

II. Related Error Measures

Three representative object-based empirical discrepancy error measures are presented in this section. The measure proposed by Martin [11] sought to quantify the consistency between segmentations with differing granularities and considered a fixed tolerance to refinement scheme. Polak [5] devised a similar approach (OCE), seeking to overcome the limitations found in Martin's measure: the lack of penalty for over- or under-segmentation. When, for example, an over-segmented image happens to be a refinement of an under-segmented image, Martin's measure would assume the two to be consistent, therefore, consider both to be similar. Finally, we describe Arbelaez method [12], which provides more consistent results when compared with OCE. Currently, Arbelaez's is the measure employed in the evaluation of image segmentation methods by the Computational Vision Group at Berkeley University [1].

A. Martin error measure

Given two input segmentations S_1 and S_2, Martin error measure produces a real-valued output in the range $[0, 1]$, where zero means no error and 1, maximum error. The main feature of this measure is its tolerance to refinement, ie, if one segment in S_1 is a subset of another segment in S_2, for example, then a pixel belonging to both segments is located in an area of refinement. Therefore, the local error should be zero. In the absence of subset relationship, then the two areas overlap inconsistently and the local error should be non-zero.

Let p_i be a pixel belonging to segments of segmentations S_1 and S_2. Let $\setminus$ denote set difference, $|x|$ the cardinality of set x and $R(S, p_i)$ the set of pixels corresponding to the region in segmentation S which contains pixel p_i. Martin defines a local refinement error as:

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \setminus R(S_2, p_i)|}{|R(S_1, p_i)|} \quad (1)$$

It is clearly non symmetric and encodes a measure of refinement in just one direction: $E(S_1, S_2, P_i)$ is zero if S_1 is a refinement of S_2, but not vice versa. This local error measure is extended to compute an error measure for the entire image: Global Consistency Error (GCE) and Local Consistency Error (LCE). The former forces all refinements to

[1] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

be in the same direction, while the latter allows refinements in distinct directions, in different parts of the image:

$$GCE(S_1, S_2) = \frac{1}{n} \min\{\sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i)\} \tag{2}$$

$$LCE(S_1, S_2) = \frac{1}{n} \sum_i \min\{E(S_1, S_2, p_i), E(S_2, S_1, p_i)\} \tag{3}$$

where n is the number of pixels in the image.

Clearly, GCE is stricter than LCE, as LCE $\leq$ GCE. These error measures tolerate under- or over-segmentation (as a consequence of their intended purpose for comparing human segmentations), unlike the OCE error measure, which will be presented below. To compensate for the non compliance to the symmetry axiom, in practice, both measures are computed and the one which yields the lowest value is selected.

B. OCE error measure

Polak et al.[5] proposed the Object-level Consistency Error (OCE) to compare two segmentations. Unlike Martin's, OCE penalises both over- and under-segmentation. Let $I_g = A_1, A_2, \cdots, A_M$ be a reference segmented image, where A_j is the jth segment in I_g. Likewise, $I_s = B_1, B_2, \cdots, B_N$ is a segmented image, where B_i is the ith segment in I_s. Let $|A|$ be the cardinality of set A. OCE initially computes a partial error measure as:

$$\begin{aligned} E_{g,s}(I_g, I_s) &= \sum_{j=1}^{M} [1 - \sum_{i-1}^{N} \frac{|A_j \cap B_i|}{|A_j \cup B_i|} \times W_{ji}] W_j, \\ W_{ij} &= \frac{\overline{\delta}(|A_j \cap B_i|)|B_i|}{\sum_{k=1}^{N} \overline{\delta}(|A_j \cap B_k|)|B_k|}, \\ W_j &= \frac{|A_j|}{\sum_{l=1}^{M} |A_l|} \end{aligned} \tag{4}$$

where $\delta(x)$ is a function that returns 1, if the input is 0; or returns 0, otherwise; and $\overline{\delta}(x) = 1 - \delta(x)$. OCE can be then defined as follows:

$$OCE(I_g, I_s) = min(E_{g,s}, E_{s,g}) \tag{5}$$

Despite the difference in notation, the terms $R(S_1, p_i)$ (Eq. 1) and A_j (Eq. 4) are equivalent. However, the key difference between OEC and Martin's measures (GCE or LCE) lies on the denominator of the error formula (Eq. 4) which employs the union of the two segments ($|A_j| \bigcup |B_i|$), as opposed to one of the segments alone (Eq. 1). Eq. 4 is based on the Jaccard index (Eq. 6) which allows penalisation of both over- and under-segmentations.

C. Alberlaez error measure

The Jaccard Index has also been used by Arbelaez [12] to create another object-based error measure. Also known as the overlap between two regions R and R', the Jaccard Index can be defined as:

$$O(R, R') = \frac{|R \cap R'|}{|R \cup R'|} \tag{6}$$

Arbelaez then defines the covering of a segmentation S by a segmentation S' as:

$$C(S' \rightarrow S) = \frac{1}{N} \sum_{R \in S} |R| \cdot \max_{R' \in S'} O(R, R') \tag{7}$$

where N represents the number of pixels in the image.

The covering of any machine-generated segmentation S by a set of reference segmentations $\{G_i\}$ can be computed as the average of the covering of S with each human segmentation G_i. Like OCE error measure, Arbelaez measure does not satisfy the symmetry axiom. As a result Arbelaez suggests two quality region descriptors: the covering of S by $\{G_i\}$ and the covering of $\{G_i\}$ by S.

III. Proposed AOM Error Measure

Let $S = R_1, R_2, \cdots, R_n$ and $S' = R'_1, R'_2, \cdots, R'_{n'}$ be two segmented images composed of n and n' regions, respectively. R_i and R_j are the ith and jth regions of S and S', respectively. An intersection matrix M of size $n \times n'$ is defined as:

$$M_{ij} = |R_i \cap R'_j|. \tag{8}$$

The error E between S and S' is given by

$$E(S, S') = 1 - \frac{1}{N} \sum_{k=1}^{min(n,n')} \max_k(M), \tag{9}$$

where N is the number of pixels in the image and $\max_k(M)$ returns the k-th largest element of M. Recursive function max is detailed in Algorithm 1.

Algorithm 1 Function **max** returns the largest intersections of matrix M

Require: Intersection matrix M and an empty vector L.
Ensure: Largest intersections of matrix M.

```
1: Function Max(M, L)
2: if Size(L) = min(n, n') then
3:     return L
4: end if
5: l ← FindLargestElement(M)
6: L.PushBack(l)
7: M ← RemoveRowColum(M, l)
8: Max(M, L)
9: End Function Max
```

Notice that, at each step of the function, both row and column, where the largest element lies, are removed (line 7 of Alg. 1), reducing the dimension of M. By doing so, we

guarantee that not only the largest intersection per object is returned by function max, but also that intersections among small objects are not ignored.

The penalty parameter α

Suppose S is a ground-truth or reference segmented image and S' is a computer-generated segmentation of the same image. Ideally, for a correct segmentation, a region $R_i \in S$ should have a corresponding similar region $R'_i \in S'$. In this case, the error E would be 0, ie, an intersection area of 100%. Over-segmentation occurs when the object $R_i \in S$ has a number s ($s \geq 2$) of corresponding segments $R' \in S'$. Our penalty criterion seeks to find s and also determine the penalty values for all the k largest intersections in M, as given by function max (Alg. 1). The penalty is computed as:

$$p_k = \begin{cases} \frac{1}{\alpha_s}, & \text{if } \alpha_s \geq 1 \\ 1, & \text{otherwise} \end{cases} \quad (10)$$

where $\alpha \in [0, 1]$ is given by the user. The parameter α dictates the weight of over-segmentation by considering only a percentage of s. However, to avoid rewarding over-segmentation, the penalty is computed only if $\alpha_s \geq 1$. Finally, we can derive, from Equation 9, a new error measure with an embedded penalty factor:

$$E_p(S, S') = 1 - \frac{1}{N} \sum_{k=1}^{min(n,n')} \max_k(M) penalty(k), \quad (11)$$

where function $penalty$ returns the corresponding over-segmentation penalty values for all k largest intersections in M. Implementation details are given in Algorithm 2.

AOM satisfies the following properties:

1) $0 \leq E(S, S') \leq 1$.
2) $E(S, S') = 0$, if $S = S'$.
3) $E(S, S') = E(S', S)$, since $\max(M) = \max(M^T)$

We illustrate the use of our error measure by referring to Figure 1. Figures 1(a) and 1(b) depict a reference segmented image S and a machine-generated segmentation S', respectively. They are both 10×10 images, with a total 100 pixels each. The error value is computed as follows:

1) Create matrix M ($M_{ij} = |R_i \cap R'_j|$),

$$M = \begin{array}{c} \\ R_1 \\ R_2 \end{array} \begin{array}{c} \begin{array}{cccc} R'_1 & R'_2 & R'_3 & R'_4 \end{array} \\ \begin{pmatrix} 80 & 0 & 0 & 0 \\ 0 & 5 & 5 & 10 \end{pmatrix} \end{array}$$

2) Compute the largest intersections in M (Algorithm 1). Since $\min(n, n') = 2$,

$$\max(M) = [80, 10]$$

3) Compute the error measure, by applying Equation 9,

$$E(S, S') = 1 - \frac{80 + 10}{100}$$
$$= 0.1$$

Algorithm 2 Funtion **Penalty** for each largest intersection.

Require: Intersection matrix M and penalty parameter α
Ensure: Penalty value for each largest intersection.

```
Function Penalty(M, α)
if Columns(M) > Rows(M) then
    return M ← Transpose(M)
end if
Zeros(L) {Create a vector of zeros}
for i = 1 to Rows(M) do
    for j = 1 to Columns(M) do
        if M(i, j) > 0 then
            L(i) ← L(i) + α
        end if
    end for
    if L(i) < 1 then
        L(i) ← 1
    end if
    L(i) ← 1/L(i)
end for
return L
End Function Penalty
```

$E(S, S') = 0.1$ means that S and S' are very similar. However, object $R_2 \in S$ was over-segmented $[R'_2, R'_3, R'_4] \in S'$, and that is prone to penalty. Another error measure could be computed as:

1) Find the penalty values for the two largest elements in M, for $\alpha = 0.5$ (Algorithm 2).

$$penalty(M) = [1, 0.666]$$

2) Compute the error, by applying Equation 11

$$E_p(S, S') = 1 - \frac{80 + 6.66}{100}$$
$$E_p(S, S') = 0.133$$

For $\alpha = 0.5$, the error between S and S' increases from 0.1 to 0.133, penalising the existing over-segmentation of object $R_2 \in S$, which corresponds to the second largest intersection in M. Notice that the value changes from 10 to 6.66.

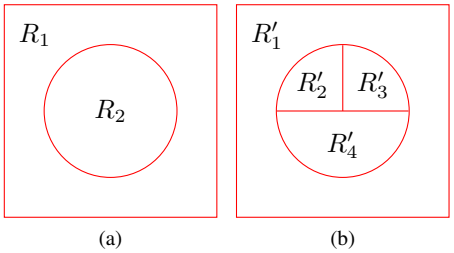


Fig. 1. Illustrating AOM error measure with synthetic images: (a) reference segmented image S with $n = 2$; (b) machine-generated segmented image S', with $n' = 4$ regions. Region sizes, in number of pixels: (a) $|R_1| = 80$, $|R_2| = 20$; (b) $|R'_1| = 80$, $|R'_2| = 5$, $|R'_3| = 5$, $|R'_4| = 10$.

IV. Experiments

We compare the proposed measure, AOM, with OCE and Arbelaez error measures. Martin's measure is not considered, as it has been outperformed by OCE measure, as shown by Polak [5]. The first experiment aims to show how the proposed measure behaves in the presence of over-segmentation. By using a set of synthetic over-segmented images, an expected theoretical error output is deduced. Performance of Arbelaez and OCE are also provided. The second experiment evaluates AOM performance and behaviour, as the penalty parameter changes. A set of over-segmentations originated from a natural scene image is employed. The third experiment shows the AOM performance for similar segmentations and the advantages over OCE and Arbelaez. Finally, we provide the computational times for the three error measures.

All three error measures have been written in C++ language. For the sake of fairness, the coding of all measures has been a direct translation of the formulas proposed by the authors, as described in this paper, with no optimization or sophisticated data structures. The code is available for download [2]. All reference or ground-truth segmented images have been taken from the Berkeley Database Segmentation Dataset (BDSD300) [11].

A. Over-segmentation Behavior

To show the behaviour of our metric while increasing over-segmentation, we created a set of synthetic images which correspond to a series of recursive segmentations. These images are illustrated in Figure 2. Given S_0 (Fig. 2.a) as a ground-truth single-region segmented image, we recursively computed six successive segmentations (Fig. 2.b to 2.g), in which each new region has half of the area of region that has been split. From this setup, an expected theoretical error can be expressed as:

$$E(S_i) = \sum_{k=0}^{n-1} \frac{1}{2^{k+1}} \tag{12}$$

where n is the number of objects in each S_i segmented image of Figure 2.

The plot of Figure 3 shows the error measures for different number of segmentations, for all three metrics. As expected, the error for segmentations S_0 and S_1 is 0.5 for all metrics, as S_1 has two identical regions. However, the three measures behave differently as the number of segmentations increases (S_2 - S_6). Since Arbelaez does not satisfy the axiom of symmetry, it has been depicted by two curves. (OCE also does not satisfy the same axiom, but the minimum of two possible values is selected). Although OCE penalises over-segmentation, it is not consistent with the increasing over-segmentation from S_2 to S_6. The error values returned from S_2 to S_6 are very similar, yielding almost a flat line. The same behaviour applies to Arbelaez measure. On the other hand, AOM reflects the penalty on over-segmentation more clearly,

[2]https://goo.gl/PkeeOP

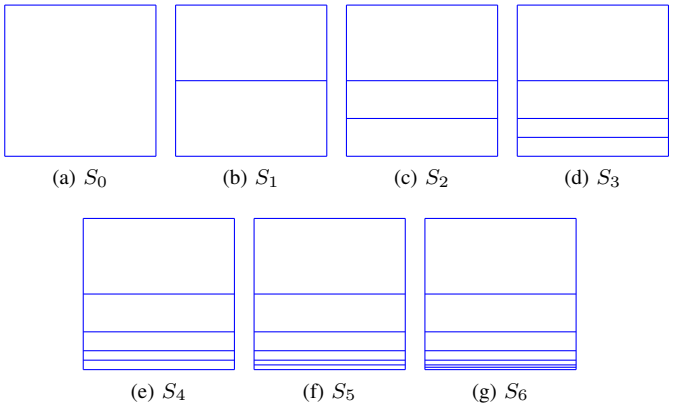


Fig. 2. Synthetic images: (a) ground-truth segmentation; (b-g) six recursively created segmentations.

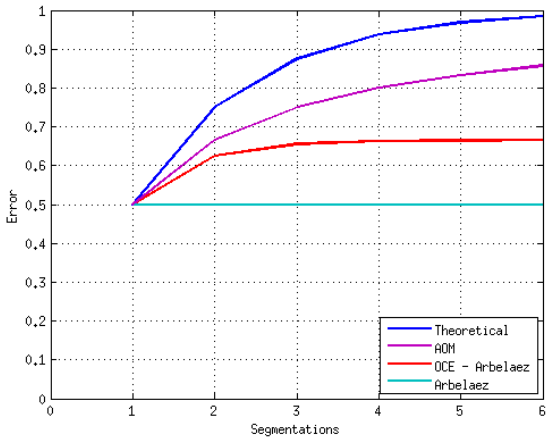


Fig. 3. Behavior of error measures on over-segmentation: AOM (magenta); OCE (red); Arbelaez (red and green) and theoretical error (blue). The proposed measure is closer to the expected theoretical error. For this synthetic set of segmentations, one of Arbelaez measures overlaps OCE, hence the same color representation.

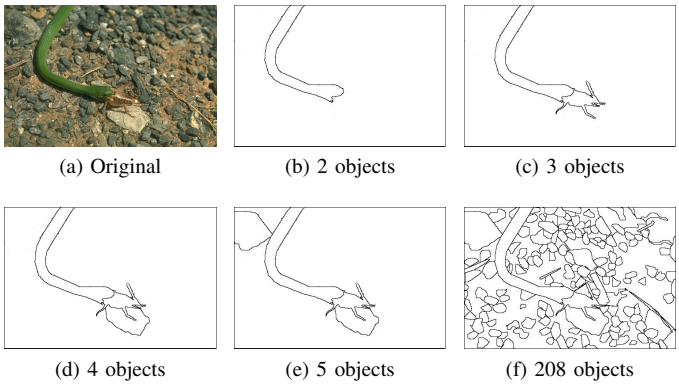
producing results closer to the expected theoretical values. In this experiment, the average value $\alpha = 0.5$ has been employed. As the value of α increases, the AOM curve gets even closer to the expected theoretical error. If the penalty parameter is switched off ($\alpha = 0$), over-segmentation is ignored, and AOM curve overlaps Arbelaez green curve, ie, error = 0.5.

B. Varying the penalty parameter

We evaluate AOM performance and behaviour, as the penalty parameter varies in the range $[0, 1]$. A natural scene image (Fig 4.a) and its human over-segmented counterpart with 208 regions (Fig 4.f) have been used. We removed object by object from Fig. 4.f, yielding decreasing over-segmented images with 193, 87, 45, 5, 4, 3 and 2 regions. Figures 4.b to .c show the images for 2, 3, 4 and 5 regions, respectively. Using the 2-region segmentation (background and snake) as a reference image, we then compute the error against all 7 remaining over-segmented images for values of α in the range

$[0, 1]$, with 0.1 increment. Figure 5 depicts all 7 curves, where the blue curve at the bottom corresponds to the error for 2-region $\times$ 3-region images and the black curve on the top, the error for 2-region $\times$ 208-region images. A barely visible yellow curve, overlapped by a black curve on the top, is also part of the plot.

Consistent error measures across all segmentations, along the entire range of α values, are captured by all 7 curves. The error remains low for two very similar segmented images (Figures 4.b and 4.c), even for large penalty values. On the other hand, when segmented images differ considerably (the top four curves, representing the error for 2-region image against 45-, 87-, 193 and 208-region images) the error is high, and remains constantly and equally high, for penalty values ≥ 0.3. However, for lower penalties, the measure is still capable of telling the difference among those segmentation.



(a) Original (b) 2 objects (c) 3 objects

(d) 4 objects (e) 5 objects (f) 208 objects

Fig. 4. Penalty parameter evaluation: (a) original image; (b-e) segmentations with increasing levels of refinement; (f) original human segmentation.

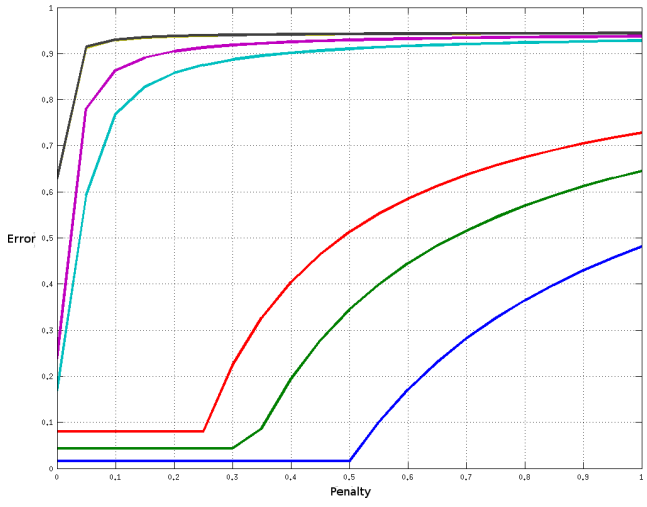
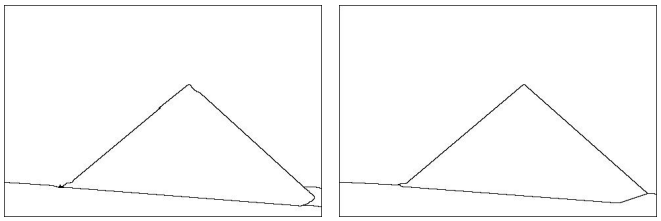


Fig. 5. Segmentation errors for the 2-region segmented image (Fig. 4.b) used as ground-truth, for different penalty values. From bottom to top: 3-, 4-, 5-, 45-, 87-, 193- and 208-region segmented images. Curves for 193- and 208-region segmentations are overlapped.

C. Similar segmentations

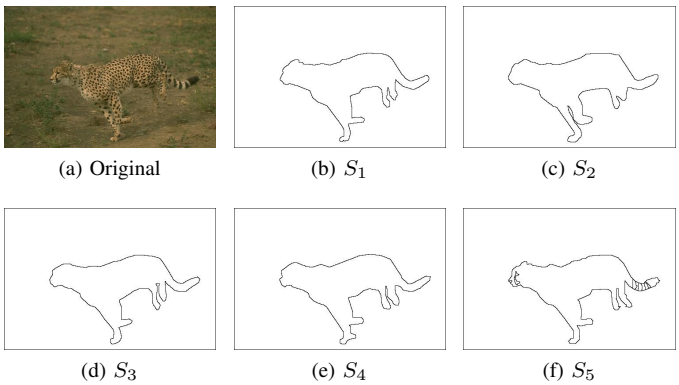
Given its subjective nature, human segmentation of natural images tend to be very diverse mitigating the analysis of metrics for over-segmented images. However, error measures must also be consistent while evaluating similar images. In this experiment, we examine AOM under this perspective. Figure 6 shows two similar human-segmented images, arbitrarily selected. Intuitively, the error between them should be near zero, as there is no apparent over-segmentation. Results obtained for AOM (with penalty switched off, $\alpha = 0$), Arbelaez and OCE were 0.0134, 0.0231 and 0.407, respectively. Notice that OCE excessively penalises over-segmentation, producing non-consistent error values. However, for applications in which minor errors should not be tolerated, we can adjust AOM penalty factor to reflect this behaviour. For $\alpha = 0.5$, for example, $AOM = 0.369$.



(a) S_1 (b) S_2

Fig. 6. Two human-segmented images with subtle differences.

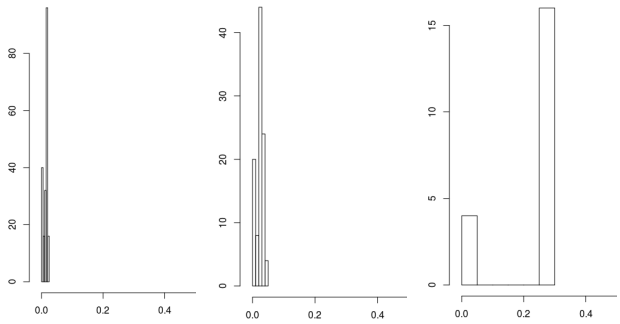
A more comprehensive example for the evaluation of similar human-segmented images is given below. Figure 7 illustrates an original image and 5 similar segmentations $(S_1, \cdots S_5)$ produced by 5 different people. Error is computed by comparing each segmentation with the 4 remaining ones, yielding a symmetric matrix of errors. Results for AOM, Arbelaez and OCE are plotted as histograms, as shown in Figure 8. Both Arbelaez and AOM (with $\alpha = 0$) exhibit consistent results, with error values close to zero. According to OCE, images are not so similar, with errors up to 0.3. The larger variance in Arbelaez, compared with AOM measure, is credited to the non-conformity of the former to the symmetry axiom.



(a) Original (b) S_1 (c) S_2

(d) S_3 (e) S_4 (f) S_5

Fig. 7. Intra-class similarities: (a) original image; (b-f) 5 similar human-segmented images produced by 5 different people



(a) AOM (b) Arbelaez (c) OCE

Fig. 8. Histograms of the 3 error measures for the 5 segmentations of Fig. 7.

Processing times

To compare the processing times for all three error measures, we randomly selected 2 segmentations of 15 images from Berkeley dataset. The mean processing times are shown in Table I. OAM is 2.777 times faster than Arbelaez and 115.906 times faster than OCE.

TABLE I
MEAN PROCESSING TIME OF THE THREE ERROR MEASURES IN SECONDS.

AOM	Arbelaez	OCE
0.164	0.456	19.03

V. CONCLUSIONS AND FUTURE WORK

In this paper we propose a new object-based empirical discrepancy error measure, named AOM, which provides a parameter for controlling the granularity segmentation level. For fairness, while comparing the proposed method with existing techniques, the parameter has been switched off and the experiments have shown that our method outperforms both OCE and Arbelaez error measures. We have shown, in this case, that AOM produces more consistent values not only for similar images, but also for images with over-segmentations. This statements holds for both synthetic and natural scenes segmentations.

It has also been shown that, by altering the parameter α, AOM consistently penalises over-segmentation. Hence, by providing means of controlling the desirable degree of refinement in segmentation, this new error measure can be employed in a broader range of applications, especially those in which the importance of over-segmentation should be sometimes ignored or highly considered.

One limitation of AOM is that it cannot be used to compare images with different dimensions. It could be argued that the introduction of a parameter renders the proposed method a generalisation problem. We believe, however, it is a pro rather than a con. As we have shown in the experiments, AOM outperforms Arbelaez and OCE for $\alpha = 0$. Therefore, in most cases the penalty parameter could simply be switched off. However, for applications with different tolerances to refinement, α could be changed and AOM would still reflect the expected error. This is not possible with Arbelaez or OCE.

As future work, we intend to demonstrate the triangle inequality axiom for the proposed measure, as well as showing its suitability for the analysis of segmentations with different scales, and how the value of the α parameter changes in relation to the image scale. We also intend to evaluate AOM performance on clustering separability.

ACKNOWLEDGMENT

The authors would like to thank FAPESP (2012/24036-1) - Brazil, for supporting this research.

REFERENCES

[1] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260 – 280, 2008.

[2] M. C. Shin, D. B. Goldgof, and K. W. Bowyer, "Comparison of edge detector performance through use in an object recognition task," *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 160–178, 2001.

[3] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

[4] J. S. Cardoso and L. Corte-Real, "Toward a generic evaluation of image segmentation," *Image Processing, IEEE Transactions on*, vol. 14, no. 11, pp. 1773–1782, 2005.

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

[6] Y. W. Lim and S. U. Lee, "On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques," *Pattern Recognition*, vol. 23, no. 9, pp. 935 – 952, 1990.

[7] M. Wollborn and R. Mech, "Refined procedure for objective evaluation of video object generation algorithms," *Doc. ISO/IEC JTC1/SC29/WG11 M*, vol. 3448, p. 1998, 1998.

[8] K. C. Strasters and J. J. Gerbrands, "Three-dimensional image segmentation using a split, merge and group approach," *Pattern Recognition Letters*, vol. 12, no. 5, pp. 307–325, 1991.

[9] J. Charles, L. I. Kuncheva, B. Wells, and I. Lim, "An evaluation measure of image segmentation based on object centres," in *Image Analysis and Recognition*. Springer, 2006, pp. 283–294.

[10] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A measure for objective evaluation of image segmentation algorithms," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 2005, pp. 34–34.

[11] 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," *ICCV*, pp. 416–423, 2001.

[12] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: an empirical evaluation," *Computer Vision and Pattern Recognition*, pp. 2294–2301, 2009.