Comparison of segmentation evaluation methods

Štěpán Šrubař
FEECS VSB-TUO
17. listopadu 15
708 33, Ostrava, CS
stepan.srubar@vsb.cz

ABSTRACT
Quality of segmentation depends on evaluation method we use. In case of specific images we use specific ways how to define quality. In general cases we have to use common methods. They differ in complexity as well as in quality. This article presents some of these methods and shows comparison of their quality on large image set.

Keywords
Segmentation, evaluation.

1 INTRODUCTION
Segments in segmentation can be measured for some specific properties (perimeter, area, curvature) or we can process or modify image information in each segment separately. For perfect results, we need high quality segmentation which is often created by a segmentation algorithm. Therefore, quality of the algorithm should be measured by an evaluation method and its quality should be evaluated as well.

2 EVALUATION METHODS
This article includes over 30 evaluation methods. Some methods needed small modification for evaluation of segmentations. Methods are divided into 6 categories which are provided in the following subsections.

2.1 Segment and Intersection Size Based Methods
Multi-class Error type I and type II by W. A. Yasnoff et al. [YMB77] use intersection of segments. Final methods SM1 and SM2 are just weighted sums. Pal and Bhandari used difference and ratio of size of segments in their Symmetric Divergence (SYD) [PB93]. D. Martin et al. [Mar02] used relative size of intersection in Global (GCE), Local (LCE) and Bidirectional Consistency Error (BCE). I propose Global Bidirectional Consistency Error (GBCE) which comes out of GCE and BCE. Larsen in L sums maxima of relative intersection sizes [LA99]. Q. Huang et al. in normalized Hamming Distance (HD) used sizes of intersections [HD95]. Van Dongen proposed metric using maximal intersections (VD) [VD00]. Meilă and Heckerman used sizes of intersections for predefined correspondence of segments (MH) [MH01]. Nearly the same was proposed later by Cardoso and Corte-Real [dSCCR05] as Partition Distance (PD).

2.2 Methods Using Distance Measure
Yasnoff et al. compute squares of distances of pixels to corresponding segments (YD) [YMB77]. Strasters and Gerbrands used the same approach in F but they normalize it by number of mis-classified pixels and a parameter [SG91]. Figure of Merit by Pratt [Pra78] uses similar evaluation expression but evaluates border pixels. Necessary natural extension is presented here as SFOM. Monteiro and Campilho combine linear and logarithmic distance (MC) [MC06]. Paumard’s Censored Hausdorff Distance was also extended (SCHD) [Pau97]. It is based on Hausdorff distance used for sets. Q. Huang and B. Dom proposed more methods but only one which uses average of distances is well defined and usable (H2avg) [HD95]. Segmentation Difference (SD) method computes discrepancy in number of segments and distance of pixels [Sru10, Sv11, Sru11]. For evaluation only distance is used. There are four alternatives of SD denoted by index.

2.3 Methods Based on Counting of Couples
If you take two random pixels, they could lie in the same segment or in different segments in the first or the second segmentation. Using combination of these properties we get four types of couples. Number of couples of each type is used in following methods: JC by Jaccard [BHEG02], FM by Fowlkes and
2.4 Methods with Statistical Approach

Yasnoff and Bacus proposed Object Count Agreement (OCA) which uses number of segments and sets of segments [YB84]. Entropy of segments is used in Normalized Mutual Information (NMI) [SGM00]. Another metric - Variation of Information (VI) was proposed by Meilă in [Mei03].

2.5 Graph Theory Methods

Method based on graph theory was proposed by Jiang et al. in [JMB06]. Segments represent nodes in graph while their correspondences are vertexes. Result is sum of all vertexes in its maximum-weight bipartite graph.

2.6 Other Methods

Last method Fragmentation (FRAG) [SG91] uses difference of number of segments. Result can be tuned by two parameters.

3 METHODOLOGY OF COMPARISON

All implemented methods are practically tested on image data set consisting of pictures taken by a typical camera [MFTM01]. They consist of sample images and their ground truth segmentations.

Comparison of two segmentations results in a number. In each comparison we know if these segmentations belong to the same image or not. According to that property results should split into two clusters (same vs. different segmentations). Typically, result of segmentations belonging to the same image are low while results from segmentations from different images are high. Therefore, we could find optimal threshold to separate results into the two categories. Still, there could be results on the wrong side of the threshold. We will divide their number by size of the whole cluster and call them false acceptance (FA) or false rejection (FR) whether they lie on the one or the other side of the threshold. The threshold can be different for each method.

4 RESULTS

Methods were implemented and measured. Symmetric methods evaluated 3138496 couples of segmentations, while each asymmetric method processed 6276992 couples of segmentations. Results are presented in figure 1. All methods reaching 50% are practically unusable. From the results we could see that local refinement (LCE, SD) is better than intolerance to refinement (BCE) which is better than tolerance to global refinement GCE. Results can also be put on the time axis (see figure 2) according to year in which their methods were published. Last graph (figure 3) shows statistical quality of each approach corresponding to categorization of the methods.

5 CONCLUSION

Segmentation evaluation methods have very various quality. Some of them are focused on a small part of information from the whole segmentation and the quality
is therefore poor. Even some recently proposed methods do not assure high quality. The best results were provided by method SD3 which uses grouping of segments and distance measuring. This quality measurement is one of the biggest according to the size of test set as well as the number of methods which can ensure high level of objectivity.

6 REFERENCES


