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 SM_1 and SM_2 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 ($H_{2\mu}$) [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

Mallows [FM83], W_I by Wallace [Wal83] and M by Mirkin [Mir96] (which is even a metric) and Rand index [Ran71]. The latter method was extended by probability of pixels (*PRI*) [UPH05].

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 [JMIB06]. 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.

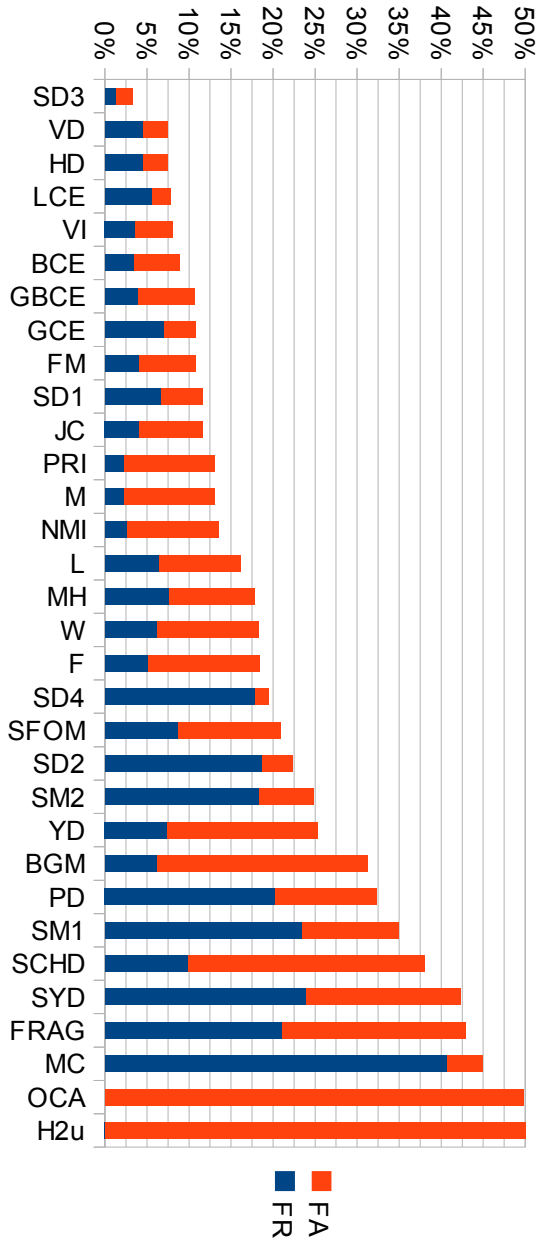

Figure 1: Results of segmentation-segmentation methods for Berkeley data set.

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

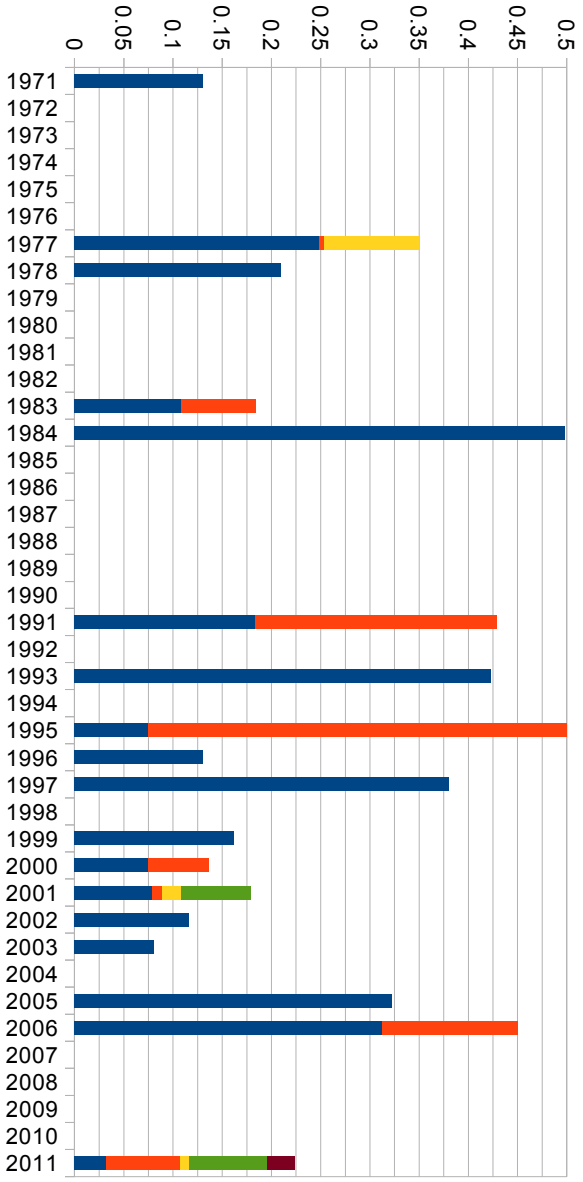

Figure 2: Review of history of methods and evolution of quality. Overall quality is shown only. Methods are stated from left to right. 1971: *PRI*, 1977: SM_2, *YD*, SM_1, 1978: *SFOM*, 1983: *FM*, *W*, 1984: *OCA*, 1991: *F*, *FRAG*, 1993: *SYD*, 1995: *HD*, $H_{2\mu}$, 1996: *M*, 1997: *SCHD*, 1999: *L*, 2000: *VD*, *NMI*, 2001: *LCE*, *BCE*, *GCE*, *MH*, 2002: *JC*, 2003: *VI*, 2005: *PD*, 2006: *BGM*, *MC*, 2011: SD_3, *GBCE*, SD_1, SD_4, SD_2.

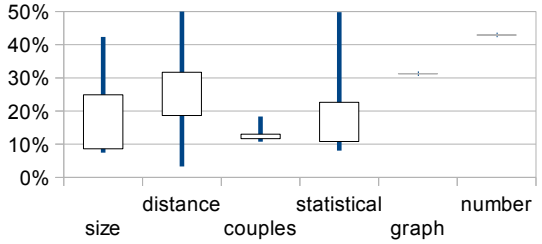

Figure 3: Overview of quality distribution in groups of methods. Minimum, maximum, lower and upper quartiles are presented.

is therefore poor. Even some recently proposed methods do not assure high quality. The best results were provided by method SD_3 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

[BHEG02] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 6–17, 2002.

[dSCCR05] Jaime dos Santos Cardoso and Luís Corte-Real. Toward a generic evaluation of image segmentation. *Image Processing, IEEE Transactions on*, 14(11):1773–1782, 2005.

[FM83] Edward B. Fowlkes and Colin L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, September 1983.

[HD95] Qian Huang and Byron Dom. Quantitative methods of evaluating image segmentation. In *Image Processing, 1995. Proceedings., International Conference on*, volume 3, pages 53–56 vol.3, 1995.

[JMIB06] Xiaoyi Jiang, Cyril Marti, Christophe Irniger, and Horst Bunke. Distance measures for image segmentation evaluation. *EURASIP J. Appl. Signal Process.*, 2006:209–209, January 2006.

[LA99] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and*

data mining, KDD '99, pages 16–22, New York, NY, USA, 1999. ACM.

[Mar02] David R. Martin. *An empirical approach to grouping and segmentation*. Phd dissertation, University of California, Berkeley, June-August 2002.

[MC06] Fernando Monteiro and Aurélio Campilho. Performance evaluation of image segmentation. In Aurélio Campilho and Mohamed Kamel, editors, *Image Analysis and Recognition*, volume 4141 of *Lecture Notes in Computer Science*, pages 248–259. Springer Berlin / Heidelberg, 2006.

[Mei03] Marina Meilă. Comparing Clusterings by the Variation of Information. *Learning Theory and Kernel Machines*, pages 173–187, 2003.

[MFTM01] David R. Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Technical Report UCB/CSD-01-1133, EECS Department, University of California, Berkeley, January 2001.

[MH01] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Mach. Learn.*, 42:9–29, January 2001.

[Mir96] Boris G. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Publishers, Dordrecht, 1996.

[Pau97] José Paumard. Robust comparison of binary images. *Pattern Recogn. Lett.*, 18:1057–1063, October 1997.

[PB93] Nikhil R. Pal and Dinabandhu Bhandari. Image thresholding: some new techniques. *Signal Process.*, 33:139–158, August 1993.

[Pra78] William K. Pratt. *Digital Image Processing*. John Wiley & Sons, Inc., New York, NY, USA, 1978.

[Ran71] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971.

[SG91] Karel C. Strasters and Jan J. Gerbrands. Three-dimensional image segmentation using a split, merge and group approach. *Pattern Recogn. Lett.*, 12:307–325, May 1991.

[SGM00] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of Similarity Measures on Web-page Clustering. In *Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI 2000), 30-31 July 2000, Austin, Texas, USA*, pages 58–64. AAAI, July 2000.

[Sru10] Stepan Srubar. Toward objective segmentation evaluation. In *International Workshops on Computer Graphics, Vision and Mathematics*, 2010.

[Sru11] Stepan Srubar. Quality measuring of segmentation evaluation methods. In *WOFEX*, 2011.

[Sv11] Stepan Srubar and Milan Šurkala. Comparison of mean shift algorithms and evaluation of their stability. In *International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2011.

[UPH05] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert. A measure for objective evaluation of image segmentation algorithms. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03*, pages 34+, Washington, DC, USA, 2005. IEEE Computer Society.

[VD00] Stijn Van Dongen. Performance criteria for graph clustering and markov cluster experiments. Technical report, National Research Institute for Mathematics and Computer Science, Amsterdam, Netherlands, 2000.

[Wal83] David L. Wallace. A Method for Comparing Two Hierarchical Clusterings: Comment. *Journal of the American Statistical Association*, 78(383(Sep., 1983)):569–576, 1983.

[YB84] William A. Yasnoff and James W. Bacus. Scene segmentation algorithm development using error measures. In *AOCH 9*, pages 45–58, 1984.

[YMB77] William A. Yasnoff, Jack K. Mui, and James W. Bacus. Error measures for scene segmentation. *Pattern Recognition*, 9(4):217 – 231, 1977.