Image Segmentation Evaluation by Techniques of Comparing Clusterings

Xiaoyi Jiang[1], Cyril Marti[2], Christophe Irniger[2], and Horst Bunke[2]

[1] Department of Computer Science, University of Münster Einsteinstrasse,
62, D-48149 Münster, Germany
xjiang@math.uni-muenster.de

[2] Institute of Informatics and Applied Mathematics, University of Bern
Neubrückstrasse, 10, CH-3012 Bern, Switzerland
{marti, iriniger, bunke}@iam.unibe.ch

Abstract. The task considered in this paper is performance evaluation of region segmentation algorithms in the ground truth (GT) based paradigm. Given a machine segmentation and a GT reference, performance measures are needed. We propose to consider the image segmentation problem as one of data clustering and, as a consequence, to use measures for comparing clusterings developed in statistics and machine learning. By doing so, we obtain a variety of performance measures which have not been used before in computer vision. In particular, some of these measures have the highly desired property of being a metric. Experimental results are reported on both synthetic and real data to validate the measures and compare them with others.

1 Introduction

Image segmentation and recognition are central problems of computer vision for which we do not yet have any solution approaching human level competence. Recognition is basically a classification task and one can empirically estimate the recognition performance (probability of misclassification) by counting classification errors on a test set. Today, reporting recognition performance on large data sets is a well accepted standard. In contrast, segmentation performance evaluation remains subjective. Typically, results on a few images are shown and the authors argue why they look good. The readers never know whether the results have been opportunistically selected or are typical examples, and how well the demonstrated performance extrapolates to larger sets of images.

The main challenge is that the question "To what extent is this segmentation correct" is much subtler than "Is this face from person x". While a huge number of segmentation algorithms have been reported, there is only little work on methodologies of segmentation performance evaluation [9]. Several segmentation tasks can be identified: edge detection, region segmentation, and detection of curvilinear structures. In this work we are concerned with region segmentation. In addition we follow the GT-based evaluation paradigm[1], in which some refer-

[1] Other paradigms include theoretical approaches, non-GT based and task-based techniques, see [9] for details.

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 344–351, 2005.

ence segmentation result (ground truth) is available and the task is to measure the difference between the machine segmentation result and the ground truth.

We propose to consider the image segmentation problem as one of data clustering and, as a consequence, to use measures for comparing clusterings developed in statistics and the machine learning community for the purpose of segmentation evaluation. We start with a short discussion of related work. Then, measures for comparing clusterings are presented, followed by their experimental validation. Finally, some discussions conclude the paper.

2 Related Work

In [5] a machine segmentation (MS) of an image is compared to the GT specification to count instances of correct detection, under-segmentation, over-segmentation, missed regions, and noise regions. These measures are based on the degree of mutual overlap required between a region in MS and a region in GT, and are controlled by a threshold T. This evaluation method is widely used for texture segmentation [2] and range image segmentation [5,7,8,12,13].

In contrast, the approach from [6] delivers one single performance measure. For each MS region R one finds the GT region R^* with the maximum intersection. Then, the total intersection between R and all GT regions other than R^* is used to compute an overall difference measure between MS and GT.

In [10] another single overall performance measure is proposed. It is designed so that if one region segmentation is a refinement of another (at different granularities), then the measure should be small or even zero. Due to its tolerance of refinement this measure is not sensible to over- and under-segmentation and may be therefore not applicable in some evaluation situations.

3 Measures for Comparing Clusterings

Given a set of objects, $O = \{o_1, \ldots, o_n\}$, a clustering of O is a set of subsets $C = \{c_1, \ldots, c_k\}$ such that $c_i \subseteq O$, $c_i \cap c_j = \emptyset$ if $i \neq j$, $\bigcup_{i=1}^{k} c_i = O$. Each c_i is called a cluster. Clustering has been extensively studied in the statistics and machine learning community. In particular, several measures have been proposed to quantify the difference between two clusterings $C_1 = \{c_{11}, \ldots, c_{1k}\}$ and $C_2 = \{c_{21}, \ldots, c_{2l}\}$ of the same set O.

If we interpret an image as a set O of pixels and a segmentation as a clustering of O, then these measures can be applied to quantify the difference between two segmentations, e.g. between MS and GT. This view of the segmentation evaluation tasks opens the door for a variety of measures which have not been used before in computer vision. As we will see later, some of the measures are even metrics, being a highly desired property which is not fulfilled by the measures discussed in the last section. In the following we present three classes of measures.

3.1 Distance of Clusterings by Counting Pairs

Given two clusterings C_1 and C_2 of a set O of objects, we consider all pairs of objects, $(o_i, o_j), i \neq j$, from $O \times O$. A pair (o_i, o_j) falls into one of the four categories

- in the same cluster under both C_1 and C_2 (The total number of such pairs is represented by N_{11})
- in different clusters under both C_1 and C_2 (N_{00})
- in the same cluster under C_1 but not C_2 (N_{10})
- in the same cluster under C_2 but not C_1 (N_{01})

Obviously, $N_{11} + N_{00} + N_{10} + N_{01} = n(n-1)/2$ holds where n is the cardinality of O.

Several distance measures, also called indices, for comparing clusterings are based on these four counts. The Rand index, originally introduced in [14], is defined as

$$\mathcal{R}(C_1, C_2) = \frac{N_{11} + N_{00}}{n(n-1)/2}$$

Fowlkes and Mallows [4] introduce the following index

$$\mathcal{F}(C_1, C_2) = \sqrt{W_1(C_1, C_2) W_2(C_1, C_2)}$$

as the geometric mean of

$$W_1(C_1, C_2) = \frac{N_{11}}{\sum_{i=1}^{k} n_i(n_i - 1)/2}, \quad W_2(C_1, C_2) = \frac{N_{11}}{\sum_{j=1}^{l} n_j(n_j - 1)/2}$$

where n_i stands for the size of the i-th element of C_1 and n_j the j-th element of C_2. The terms W_1 and W_2 represent the probability that a pair of points which are in the same cluster under C_1 are also in the same cluster under C_2 and vice versa.

Finally, the Jacard index [1] is given by

$$\mathcal{J}(C_1, C_2) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}}$$

The three indices are all similarity measures and take values out of $[0, 1]$. A straightforward transformation, e.g. $1 - \mathcal{R}(C_1, C_2)$, turns them into a distance measure such that a value of zero implies a perfect matching, i.e. two identical clusterings.

At first glance, the computation of N_{11}, N_{00}, N_{10}, and N_{01} is computationally very expensive. A naive approach would need $O(N^4)$ operations when dealing with images of size $N \times N$. Fortunately, we may make use of the confusion matrix, also called association matrix or contingency table, of C_1 and C_2. It is a $k \times l$ matrix, whose ij-th element m_{ij} represents the number of points in the intersection of c_i of C_1 and c_j of C_2, i.e. $m_{ij} = |c_i \cap c_j|$. It can be shown that

$$N_{11} = \tfrac{1}{2}(\sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij}^2 - n) \qquad N_{00} = \tfrac{1}{2}(n^2 - \sum_{i=1}^{k} n_i^2 - \sum_{j=1}^{l} n_j^2 + \sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij}^2)$$

$$N_{10} = \tfrac{1}{2}(\sum_{i=1}^{k} n_i^2 - \sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij}^2) \quad N_{01} = \tfrac{1}{2}(\sum_{j=1}^{l} n_j^2 - \sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij}^2)$$

These relationships make the indices presented above tractable for large-scale clustering problems like image segmentation.

3.2 Distance of Clusterings by Set Matching

This second class of comparison criteria is based on set cardinality alone. Using the term

$$a(C_1, C_2) \;=\; \sum_{c_i \in C_1} \max_{c_j \in C_2} |c_i \cap c_j|$$

Van Dongen [16] proposes the index

$$\mathcal{D}(C_1, C_2) \;=\; 2n - a(C_1, C_2) - a(C_2, C_1)$$

and proves that $\mathcal{D}(C_1, C_2)$ is a metric.

It can be easily shown that this index is related to the performance measure in [6]. The only difference is that the former is a distance (dissimilarity) while the latter is a similarity measure and therefore they can be mapped to each other by a simple linear transformation.

3.3 Information-Theoretic Distance of Clusterings

Mutual information MI is a well-known concept in information theory. It measures how much information about random variable Y is obtained from observing random variable X. Let X and Y be two random variables with joint probability distribution $p(x, y)$ and marginal probability functions $p(x)$ and $p(y)$. Then the mutual information of X and Y, $MI(X, Y)$, is defined as

$$MI(X, Y) \;=\; \sum_{(x,y)} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

In the context of measuring the distance of two clusterings C_1 and C_2 over a set O of objects, the discrete values of random variable X are the different clusters $c_i \in C_1$ an element of O can be assigned to. Similarly, the discrete values of Y are the different clusters $c_j \in C_2$ an object of O can be assigned to. Hence the equation above becomes

$$MI(C_1, C_2) \;=\; \sum_{c_i \in C_1} \sum_{c_j \in C_2} p(c_i, c_j) \log \frac{p(c_i, c_j)}{p(c_i)p(c_j)}$$

Here all the probability terms can be easily computed from the confusion matrix.

Note that no normalization is provided in MI. In the literature there is a normalized version of mutual information [15]

$$\mathcal{NMI}(C_1, C_2) = 1 - \frac{1}{\log(k \cdot l)} \sum_{c_i \in C_1} \sum_{c_j \in C_2} p(c_i, c_j) \log \frac{p(c_i, c_j)}{p(c_i)p(c_j)}$$

Meila [11] suggests a further alternative called variation of information

$$\mathcal{VI}(C_1, C_2) = H(C_1) + H(C_2) - 2MI(C_1, C_2)$$

where

$$H(C_1) = -\sum_{c_i \in C_1} p(c_i) \log(c_i), \quad H(C_2) = -\sum_{c_j \in C_2} p(c_j) \log(c_j)$$

represent the entropy of C_1 and C_2, respectively. This index turns out to be a metric.

4 Experimental Validation

In the following we present experiments to validate the measures defined in the last section. Some comparison work has also been done. For this purpose we consider the Hoover measure [5]. The measure from [6] was ignored because of its equivalence to the van Dongen index and the measure from [10] due to its insensitivity to over- and under-segmentation.

For the sake of clarity we consistently transformed all indices into distance measures, implying that a value of zero implies a perfect matching, i.e. two identical clusterings. Among the five performance measures from [5] we only consider the correct detection CD. The transformation $1 - \frac{CD}{\#\text{RT regions}}$ then turns it into a distance measure.

4.1 Validation on Synthetic Data

The range image sets reported in [5,13] have become popular for evaluating range image segmentation algorithms. Totally, three image sets with manually specified ground truth are available: ABW and Perceptron for planar surfaces and K2T for curved surfaces. For each GT image we constructed several synthetic MS results in the following way. A point p is selected randomly. We find the point q nearest to p which does not belong to the same region as p. Then, q is switched to the region of p provided this step will not produce additional regions. This basic operation is repeated for some $d\%$ of all points. Figure 1 shows one of the ABW GT image and two generated MS versions with different distortion levels.

One may expect that the Hoover index (correct detection) monotonically increases, i.e. becomes worse, with increasing tolerance threshold T. However, this is not true, as illustrated in Table 1 which lists the Hoover index for a

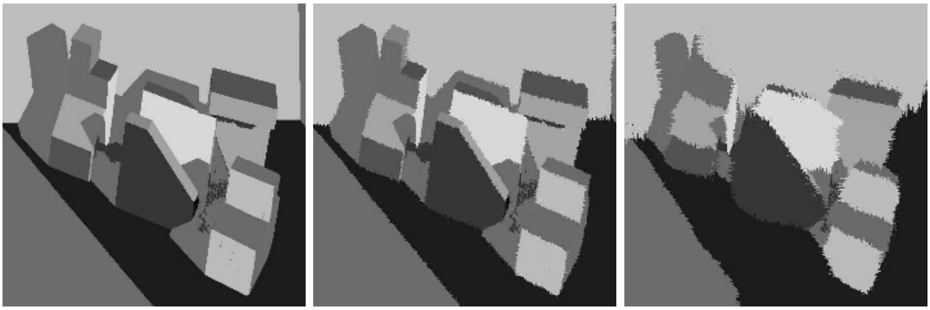

Fig. 1. An ABW image: GT (left) and two synthetic MS versions (middle: 5%, right: 50% distortion)

Table 1. Hoover index for an ABW image. The two instances of inconsistency at 20% and 30% distortion level, respectively, are underlined.

	T=0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
20% distortion	0.778	0.667	0.667	0.667	0.667	0.778	0.778	0.778	1.000	1.000
30%	0.778	0.778	0.778	0.889	0.889	0.889	0.778	0.889	1.000	1.000
40%	0.889	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

particular ABW image as a function of T and the distortion level d. There are two instances of inconsistencies. At distortion level 30%, for example, the index value 0.778 for $T = 0.85$ is lower than 0.889 for $T = 0.80$. Since we usually choose a particular value of T in practice, this kind of inconsistency may cause some unexpected effects in comparing different algorithms.

Another inherent problem of the Hoover index is its insensitivity to distortion. Basically, this index counts the number of correctly detected regions. Increasing distortion level has no influence on the count at all as far as the tolerance threshold T does not become effective. For $T = 0.85$, for instance, the Hoover index remains unchanged (0.778) at both distortion level 20% and 30%. Objectively, however, a significant difference is visible and should be reflected in the performance measures. Obviously, the Hoover index does not perform as one would expect here.

By definition the indices introduced in the last section have a high sensitivity to distortion. Table 2 lists the average values for all thirty ABW test images[2]. Obviously, no inconsistencies occur here and the values are strict monotonically increasing with a growing amount of distortion.

Experiments have also been conducted using the Perceptron image set and we observed similar behavior of the indices. So far, the K2T image set was not tested yet, but we do not expect diverging outcome.

[2] The ABW image set contains forty images and is divided into ten training images and thirty test images. Only the test images were used in our experiments.

Table 2. Comparison of synthetic MS at various distortion levels with GT: Average index values for thirty ABW test images.

	d=5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
$\mathcal{R}(C_1, C_2)$	0.024	0.041	0.055	0.068	0.080	0.091	0.102	0.111	0.120	0.129
$\mathcal{D}(C_1, C_2)$	0.027	0.046	0.063	0.078	0.092	0.105	0.117	0.128	0.138	0.149
$\mathcal{VI}(C_1, C_2)$	0.392	0.601	0.758	0.888	1.002	1.099	1.186	1.260	1.329	1.390

Table 3. Index values for thirty ABW test images

	$\mathcal{R}(C_1, C_2)$	$\mathcal{F}(C_1, C_2)$	$\mathcal{J}(C_1, C_2)$	$\mathcal{D}(C_1, C_2)$	$\mathcal{NMI}(C_1, C_2)$	$\mathcal{VI}(C_1, C_2)$	Hoover
UE	0.005	0.010	0.020	0.009	0.707	0.147	0.122
UB	0.008	0.016	0.031	0.013	0.714	0.209	0.180
USF	0.008	0.017	0.033	0.015	0.711	0.224	0.230
UW	0.009	0.017	0.033	0.019	0.848	0.236	0.435

4.2 Validation on Real Data

In [5] four segmentation algorithms have been evaluated using the Hoover measures: UE (University of Edinburgh), UB (University of Bern), USF (University of South Florida) and UW (University of Washington). Table 3 reports an evaluation of these algorithms by means of the indices introduced in this paper. The results imply a ranking of segmentation quality: UE, UB, USF, UW which coincides well with the ranking from the Hoover index (compare the Hoover index values in Table 3 and the original work [5]). Note that the comments above on Perceptron and K2T image set apply here as well.

5 Conclusions

Considering image segmentation as a task of data clustering opens the door for a variety of measures which are not known/popular in computer vision. In this paper we have presented several indices developed in the statistics and learning community. Some of them are even metrics. Experimental results have demonstrated favorable behavior of these indices compared to the Hoover measure.

Note that although experimental validation was only done in range image domain, the proposed approach is applicable in any task of segmentation performance evaluation. This includes different imaging modalities (intensity, range, etc.) and different segmentation tasks (surface patches in range images, texture regions in greylevel or color images). In addition the usefulness of these measures is not limited to evaluating different segmentation algorithms. They can also be applied to train the parameters of a single segmentation algorithm [3,12].

Given some reasonable performance measures, we are faced with the problem of choosing a particular one in an evaluation task. Here it is important to realize that the performance measures may be themselves biased in certain situations.

Instead of using a single measure we may take a collection of measures and define an overall performance measure. This way it is more likely to achieve a better behavior by avoiding the bias of the individual measures. The performance measures presented in this paper provide candidates for this approach.

References

1. A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. Proc. of Pacific Symposium on Biocomputing, 6–17, 2002.
2. K.I. Chang, K.W. Bowyer, and M. Sivagurunath. Evaluation of texture segmentation algorithms. Proc. of CVPR, 294–299, 1999.
3. L. Cingue, R. Cucciara, S. Levialdi, S. Martinez, and G. Pignalberi. Optimal range segmentation parameters through genetic algorithms. Proc. of 15th ICPR, Vol. 1, 474–477, Barcelona, 2000.
4. E.B. Fowlkes and C.L. Mallows. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78:553–569, 1983.
5. A. Hoover, G. Jean-Baptiste, X. Jiang, P.J. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher. An experimental comparison of range image segmentation algorithms. IEEE Trans. on PAMI, 18(7): 673–689, 1996.
6. Q. Huang and B. Dom. Quantitative methods of evaluating image segmentation. Proc. of ICIP, 53–56, 1995.
7. X. Jiang, K. Bowyer, Y. Morioka, S. Hiura, K. Sato, S. Inokuchi, M. Bock, C. Guerra, R.E. Loke, and J.M.H. du Buf. Some further results of experimental comparison of range image segmentation algorithms. Proc. of 15th ICPR, Vol. 4, 877–881, Barcelona, 2000.
8. X. Jiang.An adaptive contour closure algorithm and its experimental evaluation. IEEE Trans. on PAMI, 22(11): 1252–1265, 2000.
9. X. Jiang. Performance evaluation of image segmentation algorithms. In: Handbook of Pattern Recognition and Computer Vision (C.H. Chen and P.S.P. Wang, Eds.), 3rd Edition. World Scientific, 525–542, 2005.
10. D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its applications to evaluating segmentation algorithms and measuring ecological statistics. Proc. of ICCV, Vol. 2, 416–423, 2001.
11. M. Meila. Comparing clusterings by the variation of information. Proc. of 6th Annual Conference on Learning Theory, 2003.
12. J. Min, M. Powell, and K.W. Bowyer. Automated performance evaluation of range image segmentation algorithms. IEEE Trans. on SMC-B, 34(1): 263–271, 2004.
13. M.W. Powell, K.W. Bowyer, X. Jiang, and H. Bunke. Comparing curved-surface range image segmenters. Proc. of 6th ICCV, 286–291, Bombay, 1998.
14. W.M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66:846–850, 1971.
15. A. Strehl, J. Gosh, and R. Mooney. Impact of similarity measures on web-page clustering. Proc. of AAAI Workshop of Artificial Intelligence for Web Search, 58–64, 2000.
16. S. van Dongen. Performance criteria for graph clustering and Markov cluster experiments. Technical Report INS-R0012, Centrum voor Wiskunde en Informatica, 2000.