

An Information-based Measure for Grouping Quality

Erik A. Engbers¹, Michael Lindenbaum², and Arnold W. M. Smeulders¹

¹ University of Amsterdam, Amsterdam, The Netherlands,
engbers@science.uva.nl, smeulders@science.uva.nl,

² Technion — I.I.T., Haifa, Israel,
mic@cs.technion.ac.il

Abstract. We propose a method for measuring the quality of a grouping result, based on the following observation: a better grouping result provides more information about the true, unknown grouping. The amount of information is evaluated using an automatic procedure, relying on the given hypothesized grouping, which generates (homogeneity) queries about the true grouping and answers them using an oracle. The process terminates once the queries suffice to specify the true grouping. The number of queries is a measure of the hypothesis non-informativeness. A relation between the query count and the (probabilistically characterized) uncertainty of the true grouping, is established and experimentally supported. The proposed information-based quality measure is free from arbitrary choices, uniformly treats different types of grouping errors, and does not favor any algorithm. We also found that it approximates human judgment better than other methods and gives better results when used to optimize a segmentation algorithm.

1 Introduction

The performance of vision algorithms may be considered a tradeoff between their computational cost and the quality of their results. Therefore, quality measures are essential tools in the design and tuning of algorithms, in the comparison between different algorithms and in matching algorithms to tasks. As measurement tools, quality measures should be independent of the algorithms they test. They should be free of arbitrarily set parameters, provide meaningful and useful evaluations, and preferably be consistent with human judgment.

The quality of grouping algorithms, on which we focus here, may be evaluated by either task-dependent or task-independent (generic) measures. Task-dependent advocates argue that the only way to evaluate grouping quality is by considering it in the context of some application and by using the application performance as a gauge for the grouping performance. This approach is best when working on a specific application, but it does not support modular design and does not guarantee a suitable performance for other tasks [6]. In contrast, as we know, humans can consistently discriminate between good and bad segmentations. This implies that, at least in principle, task-independent measures exist [9].

Our work is done in the context of generic empirical quality evaluation, depending on a reference grouping. (This is opposed to alternatives such as *Analytic* performance evaluation [18] or empirical evaluation without a reference [10].)

Existing generic grouping quality measures rely on some kind of set difference measure which specifies a (dis-)similarity between the evaluated grouping and a reference true grouping (see e.g. [8, 16, 9]). Quality may be evaluated by, say, counting the number of incorrectly assigned pixels (additions and/or deletions), by counting the number of true groups which split or merge, or by measuring Hausdorff distances between the segments. Such measures are indeed indicative of the segmentation correctness, but the preference of one similarity measure over the other is arbitrary. One approach to addressing this confusion is to consider the tradeoff between several different measures of quality [5]. Another problem is that such similarity measures are not in complete agreement with intuitive judgment [2]. A hierarchical ground truth with multiple options substantially increases the agreement with intuition. Still, for every one of these options, the measure is still arbitrarily selected [9]. The lack of a suitable generic quality measure seems to be the main reason that subjective judgment is still the most common way for evaluating grouping results.

Unlike similarity-based approaches, we consider the grouping hypothesis as an initial guess which provides information on the unknown true grouping and reduces its uncertainty. A better grouping result provides more information about the true grouping. The amount of information may be measured in two alternative but related approaches:

Uncertainty view - Without a grouping hypothesis, the uncertainty about the true grouping is high and the number of possible true groupings is large. Knowing the grouping hypothesis reduces the number of correct grouping possibilities, or at least reduces the likelihood of some of them. Quantification of the uncertainty reduction is thus a measure of quality.

Effort view - Suppose the true group is specified by a sequence of basic measurements, such as the cues used by grouping algorithms. The length of the minimal sequence is the effort required for specifying the correct grouping. A given grouping reveals information about the unknown correct grouping and reduces the effort. Quantification of effort reduction is a measure of the hypothesis quality.

These two ways for evaluating the quality are related. When the uncertainty is large, more effort is needed to find the correct hypothesis. In information theory, a complete probabilistic characterization allows us to specify a tight relation between the code length (effort) and the entropy (uncertainty). Here, the relation exists but is not as tight. Therefore, we emphasize the effort based approach which provides a practical and intuitive procedure for grouping quality evaluation.

Similar relations between ‘effort’ and ‘quality’ form the basis of quality measures in other domains such as the earth movers distance [13] and the string edit distance [11]. In the context of image segmentation, this relation is explicit in

[15], where segmentation results are evaluated by measuring the effort it takes to manually edit the hypothesis into the correct solution. The resulting evaluation measure is the weighted sum of the performed actions in the editing phase. Our proposed method may be considered to be an automatic version of this subjective, manual approach.

We also propose a probabilistic characterization of the grouping process and show that considering the given grouping result to be a random variable gives a precise meaning to the uncertainty of the true grouping, using terms such as surprise and entropy, as defined in information theory [3]. We show how these terms are related to the quantification of quality using effort.

The proposed approach has the following advantages:

General, uniform and fair - The measure uniformly deals with various types of grouping mistakes, does not involve ad-hoc decisions or parameters and is not biased towards any particular method.

Consistent with HVS - The measure is more consistent with human judgment than other methods.

Meaningful - The measure may be interpreted by known statistical terms (from information theory).

Useful - The quality measure is practical, useful, and allows, for example the optimization of the parameters of a grouping algorithm, in a better way.

After some necessary definitions (Section 2), the proposed quality measure is presented in Section 3. A link to the uncertainty view is made in Section 4. Section 5 reports on some experiments, including psychophysics. The paper is concluded with some observations and directions for future work in Section 6.

2 Preliminaries

Let $S = \{s_1, s_2, \dots, s_n\}$ be a set of elements, and consider a grouping C of S as the partition of S into disjoint subsets, $C = \{X_1, X_2, \dots, X_m\}$, with $X_i \subseteq S$, $\cup_i X_i = S$, $X_i \cap X_j = \emptyset$ for $i \neq j$. The set of all possible groupings of set S is defined as $\mathcal{C} = \{C | C \text{ is a grouping of } S\}$. A useful grouping may not be disjoint or even unique. Here, however, we take a simplified approach and further assume that there is only one correct grouping, denoted $C_T \in \mathcal{C}$. It is straightforward to generalize the proposed measure to handle non-unique correct groupings.

A grouping algorithm provides a grouping hypothesis C_H , which is a partition as well. Usually it will not be identical to the true grouping C_T . The goal of this paper is to propose a measure, denoted $Q_T(C_H)$, for the quality of the grouping hypothesis C_H relative to a known true grouping C_T .

3 Judging grouping quality from a questions game

To quantify the effort required to get the correct grouping from the grouping hypothesis, we consider a generic procedure that asks questions about the true

unknown grouping. These questions are answered by an oracle, which relies on the true grouping and provides only “Yes/No” type answers. After a sufficient number of questions are asked, the true grouping may be inferred. The questioning procedure builds on the grouping hypothesis, and the grouping hypothesis is considered better if the number of questions, or the *effort*, is lower.

Note that this view of quality is closely related to the parlor game of “Twenty Questions” [14], where one needs to ask a minimal number of questions in order to identify a secret object. (See also [12] where such a query mechanism is used for visual object recognition.) In this context, the value of a ‘hint’ may be measured by the number of questions it can save. Correspondingly, the grouping hypothesis is considered a hint of the true grouping, and its quality is the number of queries it saves.

Thus, to evaluate the quality of C_H , the following needs to be specified:

- an oracle that knows the correct grouping
- a systematic questioning strategy R , and
- a set of questions from which a subset is drawn according to the questioning strategy.

3.1 Homogeneity Queries

The type of questions, or queries, we allow are *homogeneity queries* [17, 8]. That is, every query specifies a subset of the image and asks whether all picture elements in it belong to the same group. The oracle, knowing the correct grouping, can easily answer such questions. Adopting another query type (e.g. a Boolean function over several homogeneity queries) could lower the number of queries, but we conjecture that it would not change the relative number substantially, implying that it would not be better for *comparing* grouping hypotheses. Our experiments (Section 5.3) support this argument.

3.2 Questioning strategies

A questioning strategy suitable for the proposed quality evaluation should have two main properties. It should not be biased toward specific types of grouping results and should be efficient in the sense of not asking more questions than necessary. An optimal strategy, asking the minimal (average) number of questions would be best, and could be designed, at least in principle, from a probabilistic model (described in the next section).

Here, however, we have chosen a non-optimal strategy which is based on a split-and-merge search for the true grouping (see Section 5). We conjecture that such a strategy provides query counts which are proportional to those achieved with optimal strategies, and thus is good enough for estimating relative qualities. The experimental results, described in Section 5.3 support this conjecture.

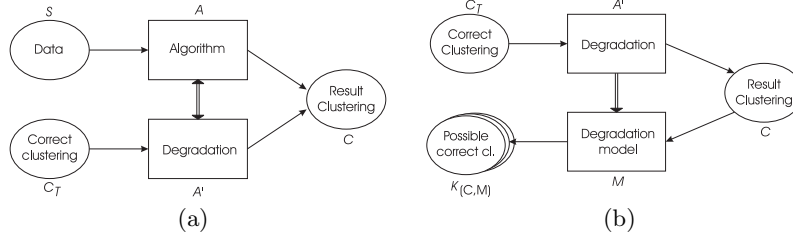


Fig. 1. Two alternative views of a grouping (clustering) process: as an algorithm that labels the data elements or as a degradation process from the true grouping to the grouping result (a), and construction of the set of all possible correct grouping (clusterings) given C and M (b).

3.3 A normalized quality measure

Let $N_q(C_H, C_T)$ be the number of queries required to know C_T from the hypothesis C_H . One possible normalized quality measure is

$$Q_T(C_H) = \frac{N_q(C_W, C_T) - N_q(C_H, C_T)}{N_q(C_W, C_T) - N_q(C_T, C_T)}, \quad (1)$$

which is maximal and 1 for the best (true) hypothesis ($C_H = C_T$) and is minimal and zero for the worst hypothesis C_W , specified as the one requiring the maximal number of questions. While this normalization is intuitive, other normalizations are possible, and may even be preferable; see [4]. In this paper we focus on the raw quantity $N_q(C_H, C_T)$.

4 A statistical notion of grouping quality

4.1 A statistical model for grouping

Grouping algorithms significantly differ in the processes they use. A quality measure, as any objective measurement tool, needs to consider only the results of the algorithms in a uniform way, using a common language, unrelated to the process carried out by the algorithm. Therefore, we consider the grouping hypothesis, provided by an algorithm A , to *also* be a result of an equivalent process, called *degradation*, denoted by $A' : \mathcal{C} \rightarrow \mathcal{C}$. This process, operating on groupings (and not on images) receives the correct grouping C_T as an input and provides that same hypothesis $C_H = A'(C_T)$ that the algorithm A delivers (see Figure 1)

The degradation process takes into account both the algorithm and the image given to it, as both of them influence the grouping hypothesis. It may be modeled stochastically as follows:

Stochastic Degradation Model:

A degradation process A' from C_T to $C = A'(C_T)$ is an instance of a random variable M drawn using some probability distribution $P_M(A')$.

The random variable M is denoted a *degradation model*. An instance of this random variable is a particular *degradation process*, which is equivalent to the action of a particular grouping algorithm on a particular image; see a more detailed modeling in [4].

4.2 The posterior distribution of groupings

For a given grouping hypothesis C_H , true grouping C_T , degradation model M , and some prior distribution $P^*(C)$ over the set of true groupings, the set of all possible correct groupings, $K_{(C_H, M)}$, which are consistent with the grouping hypothesis C_H , can be constructed as (see Figure 1):

$$K_{(C_H, M)} = \{C | C \in \mathcal{C}, P_M(A') > 0, A'(C_T) = C_H\}. \quad (2)$$

Taking a Bayesian approach, the posterior probability that a particular grouping $C \in K_{(C_H, M)}$ is the true grouping, is

$$P_{(C_H, M)}(C) = \frac{\sum_{\{A' | A'(C) = C_H\}} P_M(A') P^*(C)}{\sum_{C' \in K_{(C_H, M)}} \sum_{\{A' | A'(C') = C_H\}} P_M(A') P^*(C')}. \quad (3)$$

4.3 The quality of a grouping hypothesis as a surprise

Given the posterior probability distribution on the possible true groupings, one can construct an optimal questioning strategy, following the corresponding Huffman code [3]. It is well known that the average length of the Huffman code converges (from above) to the entropy and is thus minimal³.

The following related result is even more useful here: the number of bits required to code a message, associated with probability p , using a Huffman code, is never larger than $\lceil -\log_2 p \rceil$ [3]. The quantity $-\log_2 p$ is sometimes called *surprise* [7], in accordance with intuition: a rare message, associated with a small probability p , makes a large surprise when it appears.

If the query strategy is designed according to the Huffman code, it follows that the number of queries is not higher than the surprise associated with the event (of probability $P_{(C_H, M)}(C = C_T)$) that, given C_H , C_T is the true grouping. The quality measure Q_T (1) is monotonic in the query count. Therefore,

Claim 1 (Hypothesis Quality and Surprise) *Under the probabilistic degradation model, and with an optimal questioning strategy and unlimited queries, the proposed hypothesis quality measure Q_T is a monotonically non-increasing function of $-\log_2 P_{(C_H, M)}(C = C_T)$ - the surprise associated with this grouping.*

This relation attaches a new meaning to the proposed quality measure. A larger number of queries implies that the true grouping is more “surprising” (in the information theory sense), meaning that the hypothesis is less informative and worse.

³ $H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$ is the entropy of the random variable X .

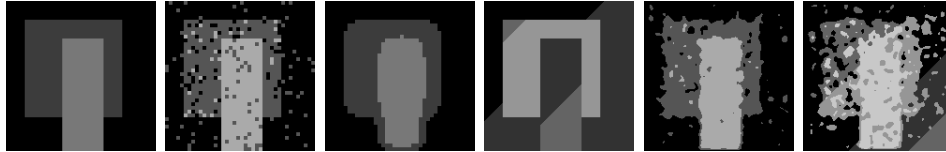


Fig. 2. The original true grouping, the three types of basic degradation: noise type, attached type, and split/merge type, and two examples of mixture degradation: noise+attached, and noise+attached+split (left to right).

A subtle issue is the choice of the degradation model and the corresponding questioning strategy. A model adapted to a particular algorithm assigns a smaller surprise to grouping errors which are typical to it and thus introduces bias. Therefore, when comparing different hypotheses without knowledge of the degradation models, we choose a degradation which characterizes the ensemble to grouping algorithms.

4.4 The quality of a grouping algorithm as (average) entropy

So far, we considered the quality of a single grouping hypothesis C_H . To evaluate an algorithm, we propose to average the number of queries required for a large number of grouping tasks. In our model, this average number converges to the average entropy

$$\overline{H(A)} = \sum_{C \in \mathcal{C}} H(P_{(C,M)}) \text{Prob}(C_H = C), \quad (4)$$

which depends only on the degradation model M and on the prior distribution of true groupings. For a detailed discussion see [4].

The elegant relation between effort and surprise (and between average effort and average entropy) holds rigorously only in the ideal case. When examining a grouping hypothesis we usually do not know the degradation model, and consequently, cannot design the optimal Huffman strategy. Moreover, the *homogeneity queries* we propose are weaker than the arbitrary queries required to specify the Huffman tree. While we do not have a provable relation, we conjecture that the query count is monotonic in the surprise and use the experimental effort as a quality measure. Some experiments, described in Section 5.3, show that this conjecture is a justified approximation.

5 Experiments

The experiments described below illustrate how the measure quantifies common types of grouping errors, show its improved agreement with human judgment and a relation between measured quality and entropy estimates, and demonstrate the proposed measure utility for optimizing a simple grouping algorithm.

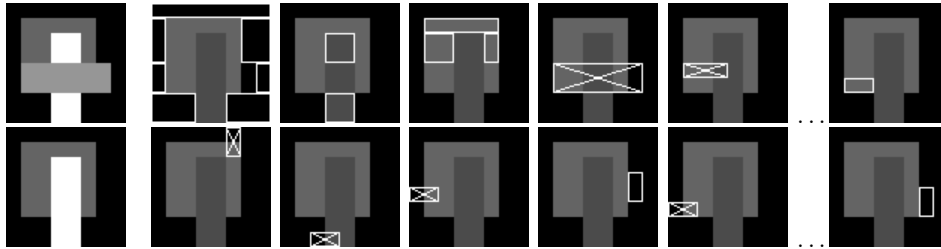


Fig. 3. An illustration of a query sequence: the two images on the left show the hypothesis (top) and the true grouping (bottom). The rest of the images describe typical queries starting from the first (top - second left). The top row shows queries from the split stage. The first three queries are done on a homogeneous region. The fourth is done on a non-homogeneous region and consequently this region is split. Queries associated with the merge stage are described in the bottom line. In every image, the white boundary rectangle (or union of such rectangles) marks the region tested for homogeneity, and a cross over this region implies that the homogeneity test failed.

5.1 Some illustrative quality measurements

Data The grouping errors considered in the first set of experiments are mixtures of three basic degradation types, illustrated in Figure 2: *noise type (independent) errors*, characterized by isolated “islands” of incorrect grouping and rough group boundaries, *attached errors*, where the errors are near the real groups boundaries and have relatively smooth shapes, and *split-and-merge errors* where large parts of the original groups are split and merged. Using synthetic grouping error generators allows us to examine the quality measures for the different types of grouping errors. Moreover, this way, the measure is not developed in the context of a particular algorithm and is not biased towards its properties.

The questioning strategy A sequence of homogeneity queries, based on the *given grouping hypothesis*, is used to find the *true grouping*. The strategy recursively splits the groups specified by the hypothesis until the subparts are homogeneous. This is done hierarchically, according to a binary regions tree, which is built for every hypothesized group. Then, the unions of the subparts are tested, so that no two (adjacent) parts which belong to the same group remain separate. See [4] for a more detailed and formal description and Figure 3 for an illustration.

The quality calculator procedure is available to the community via a web site www.cs.technion.ac.il/Labs/Is1/Research/grouping-quality/gq.htm allowing the uploading of the user’s images.

The query count measure for a variety of grouping degradations For every tested grouping, we consider two quality measures: the one we propose, and a reference one, denoted the difference measure, which is simply the minimal number of hypothesis pixels that should be corrected to get the true grouping.



Fig. 4. The dependency of the query count on the error type. The five segmentations above differ in the source of the errors. The leftmost image gets all its errors from splits and merges, while as we move to the right, the influence of attached errors and noise is stronger. The number of queries required for these hypotheses are 51, 623, 593, 914 and 1368. (Corresponding quality values (Q_T) are 0.987, 0.87, 0.84, 0.76, and 0.64.) The corresponding difference measures are 10601, 10243, 10001, 9946 and 9916. Note that even though the difference measure stays the same or even decreases, the number of queries significantly increases.



Fig. 5. The query count measures for some hand segmented images. The left-most image is assumed to be “the true one” and the rest are the other hand segmentations ordered (left to right) by increasing query count (317, 406, 548, 566 and 678).

First we considered hypotheses, degraded by different amounts of noise-based errors. Naturally, more erroneous hypotheses required more queries and have higher difference measures. Next we addressed a more complex degradation, which is a mixture of attached type and noise type errors. Here the number of queries significantly decreases when the attached type error is more dominant, even though the difference measure stays the same. The measures are markedly different when split/merge grouping mistakes are dominant. Then, in agreement with intuition, the proposed measure does not penalize every pixel of the split part as if the grouping of these pixels were completely incorrect. The number of queries required to restore a grouping dominated by this type of grouping errors is much lower than the number required for noise or attached type grouping errors; see Figure 4 for an example, and [4] for details regarding all experiments.

5.2 Comparison of hand segmented images

While we still think that the characteristics of the proposed measure are best revealed using controllable grouping errors, we tested the measure using examples from the Berkley database. We took several segmentations of the same image, chose one of them as the “true one” and examined the information revealed on it from the others. The query count may become more dependent on the small scale errors. To avoid that (as done say, in [9]), we treated thin regions of non-homogeneity as “don’t care” in the homogeneity query. See Figure 5 for an example demonstrating the good agreement with human judgment.

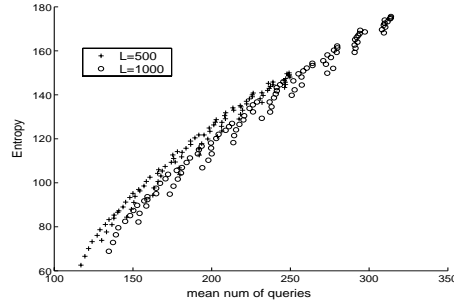


Fig. 6. The plot describes the combinatorially calculated entropy against the average query count, for two image sizes ($L = 500, 100$), $k = 30$ true groups, and a variety of degradations. (All combinations of s splits and m merges where $s, m \in [6, 8, 10, \dots, 24]$).

5.3 Average query count vs. entropy experiment

The validity of the probabilistic interpretation is tested by relating the average query count to the entropy. Picking a relatively simple example allows us to analytically approximate the entropy. We consider 1D “images” of length L . These images are truly divided into several (k) true groups, but the given hypotheses result from a degradation process in the form of m merges and s splits. (The splits do not split between true groups and the merges do not merge parts which were split in the degradation). Given such an hypothesis, the number of feasible true groupings is combinatorially calculated to be $\binom{k+s-m-1}{s} \binom{L-k-s+m}{m}$, and its logarithm is an upper bound and a good approximation of the entropy.

A 1D variant of the questioning strategy, unaware of the parameters k, m and s , is used to find the query count measure. The average number of queries is estimated for every parameter set L, k, m, s , from several randomizations of the true groupings and degradations corresponding to this parameter set.

The combinatorially calculated entropy is plotted, in Figure 6, against the average query count. The different selections for s and m specify a wide variety of different grouping errors. Still, the relation between the query count and the entropy is almost linear, and independent of the error type.

Recalling that the homogeneity queries are weaker than general queries, this result is consistent with our interpretation, and in particular, with our claim that in the ideal case the average query count is the (average) entropy. It also supports, more weakly, the claim that for a single hypothesis, the query count is proportional to the surprise. An important practical implication is that since the number of queries required using our procedure is approximately monotonic and linear in the number of queries required by the ideal strategy for a variety of grouping errors, it is an unbiased measure for comparing grouping results.

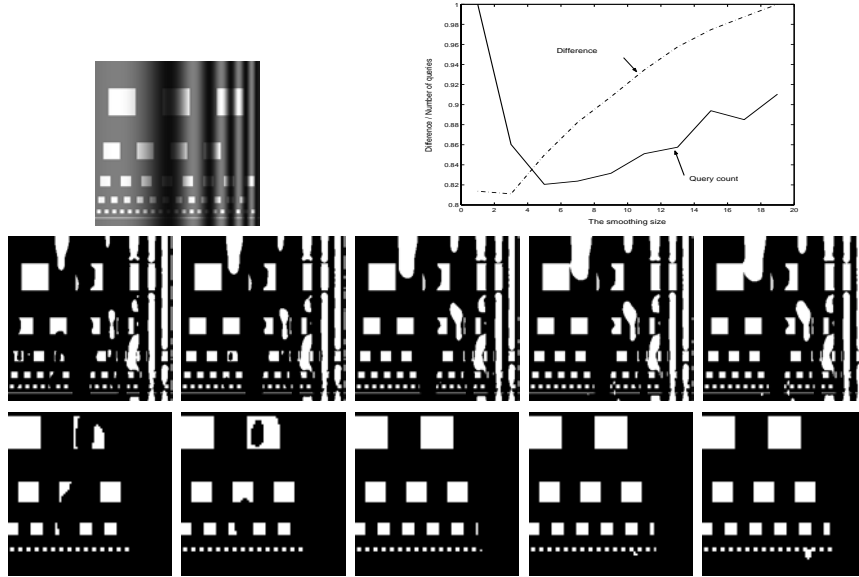


Fig. 7. The given image (top left), 5 binarized images (grouping hypotheses), corresponding to smoothing widths of 1, 3, 5, 7 and 9 (left to right, middle and bottom (zoomed version)), and the query count and the difference measures plotted against the smoothing kernel width (top right, the measures are normalized). The binarization specified by the minimal query count is indeed visually optimal.

5.4 An application - algorithm tuning

To show its utility, we used the new measure to tune a particular segmentation (binarization) algorithm. This algorithm smooths a given image, extracts edge points and uses them to set a binarization threshold surface [1]. The degree of smoothing is a free parameter of the algorithm which we optimized; see Figure 7 for the results. Note that minimizing the query count (minimal at smoothing width = 5) leads to the visually most pleasing result, while relying on the difference measure (minimal at smoothing width = 3) does not.

5.5 Psychophysics

We tested the consistency of the proposed measure with human judgment in a simple psychophysical experiment. A subject was shown a sequence of slides, each containing two pairs of grouping images, and was asked to tell which pair shows a more similar presence of objects. The pairs were of very similar groupings, one different from the other by a mixture of attached type and noise type errors. The difference measure was 1-2% for all pairs.

An answer was considered correct if it was in agreement with the query count based preference. The number of incorrect answers (an average over five

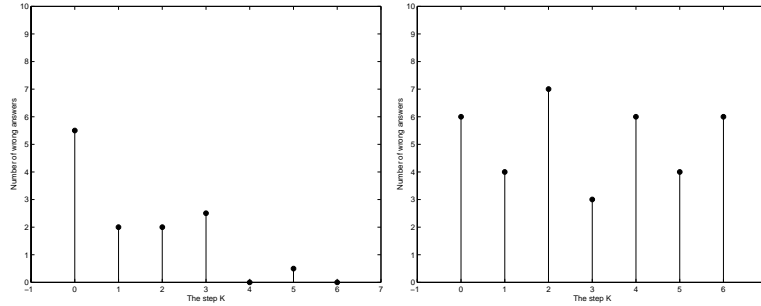


Fig. 8. The rate of incorrect answers as a function of ranking according to query based measure (left) and to additive difference based measure (right)

subjects) is plotted against the difference in quantized query counts (Figure 8). (The query counts of all pairs were quantized into 7, equally populated levels. For every difference value, 10 slides were presented.) The error rate sharply decreases for a higher difference in query count, showing an excellent agreement between the query count based measure and human judgment. Repeating this experiment for the difference measure shows little, if not zero, agreement (Figure 8).

6 Conclusions

We proposed a generic grouping quality measure, quantifying the information available from the grouping result on the true grouping. A (simple and automatic) simulated interaction between a questions strategy and an oracle was used to estimate this information. We found that the proposed measure more closely approximates human judgment than other methods and as such gives better results when used to optimize a segmentation algorithm. The proposed methods is associated with the following two main advantages:

Generality and fairness - Most previous, similarity-based measures, involve unavoidable arbitrary choices. The proposed information-based quality measure is free from such arbitrary choices, treats different types of grouping errors in a uniform way and does not favor any algorithm.

Non-heuristic justification - The number of queries is interpreted as a *surprise* in an information theory context. While the questioning strategy is not ideal, the query count was found to be approximately monotonic in the entropy, independent of the grouping error type, indicating both that this interpretation is valid and that the query count is an adequate unbiased means for comparing grouping results.

This work was done in the context of a unique ground truth. One future direction would be to generalize our measure to multiple ground truths (as was shown to be more meaningful in [9]). This could be done by finding the query count for all ground truths and calculating the quality from the minimal value.

Acknowledgments

M.L. would like to thank Ronny Roth and Neri Merhav for discussions on information theory and Leonid Glykin for his help in the experiments.

References

1. I. Blayvas, A. Bruckstein, and R. Kimmel. Efficient computation of adaptive threshold surfaces for image binarization. In *CVPR01*, pages I:737–742, 2001.
2. A. Cavallaro, E.D. Gelasca, and T. Ebrahimi. Objective evaluation of segmentation quality using spatio-temporal context. In *ICIP02*, pages III: 301–304, 2002.
3. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. Wiley-Interscience, 1991.
4. E.A. Engbers, M. Lindenbaum, and A.W.M. Smeulders. An information based measure for grouping quality. Technical Report CIS-2003-04, CS dept., Technion, 2003.
5. M. Everingham, H. Muller, and B. Thomas. Evaluating image segmentation algorithms using the pareto front. In *ECCV02*, page IV: 34 ff., 2002.
6. W. Foerstner. 10 pros and cons against performance characterization of vision algorithms. In *Performance Characteristics of Vision Algorithms*, Cambridge, 1996.
7. G.D. Forney. Information theory. unpublished lecture notes, EE dept. Stanford university.
8. D.G. Lowe. *Perceptual Organisation and Visual Recognition*. Kluwer Academic Publishers, 1985.
9. D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV01*, pages II: 416–423, 2001.
10. P. Meer, B. Matei, and K. Cho. Input guided performance evaluation. In R. Klette, H.S. Stiehl, M. Viergever, and K.L. Vincken, editors, *Performance Characterization in Computer Vision*, pages 115–124. Kluwer, Amsterdam, 2000.
11. S. V. Rice, Horst Bunke, and T. A. Nartker. Classes of cost functions for string edit distance. *Algorithmica*, 18(2):271–280, 1997.
12. W. Richards and A. Bobick. Playing twenty questions with nature. In Z. Pylishin, editor, *Computational Processes in Computer Vision: An interdisciplinary Perspective*, chapter 1, pages 3–26. Ablex, Norwood, 1988.
13. Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Proc. 6th ICCV IEEE Int. Conf. on Computer Vision*, pages 59–66, 1998.
14. C.E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30:50–64, 1951.
15. K.L. Vincken, A.S.E. Koster, C.N. De Graaf, and M.A. Viergever. Model-based evaluation of image segmentation methods. In *Performance Characterization in Computer Vision*, pages 299–311. Kluwer Academic Publishers, 2000.
16. L. Williams and K.Thornber. A comparison of measures for detecting natural shapes in cluttered backgrounds. In *ECCV98*, 1998.
17. A. Witkin and J. Tenenbaum. On the role of structure in vision. In J. Beck, B. Hope, and A. Rozenfeld, editors, *Human and Machine Vision*, pages 481–543. Academic Press, 1983.
18. Y.J. Zhang. A survey on evaluation methods for image segmentation. *PR*, 29(8):1335–1346, August 1996.