How to Complete Performance Graphs in Content-Based Image Retrieval: Add Generality and Normalize Scope

Dionysius P. Huijsmans and Nicu Sebe, *Member*, *IEEE*

Abstract—The performance of a Content-Based Image Retrieval (CBIR) system, presented in the form of Precision-Recall or Precision-Scope graphs, offers an incomplete overview of the system under study: The influence of the irrelevant items (embedding) is obscured. In this paper, we propose a comprehensive and well-normalized description of the ranking performance compared to the performance of an Ideal Retrieval System defined by ground-truth for a large number of predefined queries. We advocate normalization with respect to relevant class size and restriction to specific normalized scope values (the number of retrieved items). We also propose new three and two-dimensional performance graphs for total recall studies in a range of embeddings.

Index Terms—Multimedia information systems, information retrieval, content-based image retrieval, performance evaluation.

1 The Need for Reevaluation of CBIR Performance Characterization

The motivation for this work came up during the experiments carried out on our Leiden 19th-Century Portrait Database (LCPD). This database presently contains 21,094 photographic portraits with as many studio logos (on the separately scanned backsides). These logos are manually grouped into 1,856 nonoverlapping logo classes with an average size of almost eight relevant unordered items; relevant class size is in the range [2, 308]. A more extensive description of this ground-truth test set can be found in [8].

Each relevant class item was taken in turn to perform a content-based query by example search. Evaluations were based on the average number of its remaining relevant class members retrieved within various scopes (number of retrieved items), using different indexing feature vectors, different similarity measures, and for different relevant class sizes. By changing the size of the embedding (the number of irrelevant items in the database) for a specific relevant class size, we obtained a series of Precision-Recall (PR) curves (Fig. 1) for the performance of a ranking method based on the gray-level histogram as a feature vector and L_1 as a metric for similarity. One observes by looking at this figure that it contains both well-performing curves (the ones at the top) and bad-performing curves (the ones at the lower left side). The reason behind this widely varying performance behavior is the effect of the changing fraction of relevant items in the database. This relevant fraction, known as *generality*, is a major parameter for performance characterization that is often neglected or ignored. The fact that *generality* for a class of relevant items in a large embedding database is often ≈ 0 does not mean that its exact low level will not count. A continually growing embedding of irrelevant items around a constant size class of relevant items will normally (in the case of polysemantic or noisy keys) lower the overall PR curve (for the user) to unacceptable low levels, as is shown in Fig. 1.

In general, the dynamic growth of the database might result in a relative growth that is equal for both relevant items and irrelevant embedding items. In this case, the PR graph would remain at the same generality level. Moreover, a constant retrieval recall rate would mean that the scope, used to retrieve these relevant items, would have to increase with the same percentage as well. Hence, it would be advantageous to couple the scope to the expected size of the relevant class.

The performance characterization of content-based image and audio retrieval often borrows from performance figures developed over the past 30 years for probabilistic text retrieval. Landmarks in the text retrieval field are the books by Salton [5] and van Rijsbergen [18], as well as the proceedings of the annual ACM SIGIR and NIST TREC conferences.

Shortcomings of Single Measures. In probabilistic text retrieval, like in [18], the NIST TREC [20] and MPEG-7 descriptor performance evaluation [9] authors often go for single measure performance characterizations. These single measures are based on ranking results within a limited scope without taking into account both the size of the relevant class and the effect of changing either the size or the nature of the embedding irrelevant items. By their nature, these single measures have limited use because their value will only have a meaning for standardized comparisons, where most of the retrieval parameters, such as the embedding, relevant class size, and scope are kept constant.

Shortcomings of Precision-Recall Graphs. In the area of probabilistic retrieval, the results of performance measurements are often presented in the form of Precision-Recall and Precision-Scope graphs. Each of these standard performance graphs provides the user with incomplete

- *D.P. Huijsmans is with the Leiden Institute of Advanced Computer Science, Leiden University, PO Box 9512, 2300 RA Leiden, The Netherlands. E-mail: huijsman@liacs.nl.*
- *N. Sebe is with the Faculty of Science, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands. E-mail: nicu@science.uva.nl.*

Manuscript received 5 Dec. 2003; revised 16 Apr. 2004; accepted 9 June 2004. Recommended for acceptance by C. Schmid.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0399-1203.

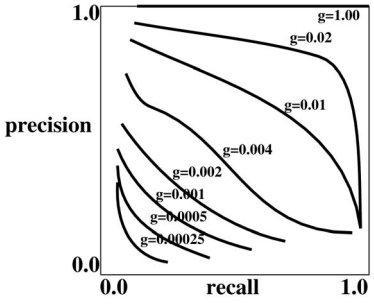

Fig. 1. Typical precision-recall curves for retrieval of a constant size relevant class of eight items embedded in a growing number of irrelevant items (max 32,000) using a ranking method based on gray-level histogram intersection. The relevant fraction or generality for each curve is given by its g value.

information about how the Information Retrieval System will perform for various relevant class sizes and various embedding sizes. *Generality* (influence of the relevant fraction) as a system parameter hardly seems to play a role in performance analysis [1], [12], [19]. Although *generality* may be left out as a performance indicator when competing methods are tested under constant generality conditions, it appears to be neglected even in cases where *generality* is widely varying (a wide range of relevant class sizes in one specific database is the most frequently encountered example).

The lack of generality information in Precision-Recall and Precision-Scope graphs makes it difficult to compare different sized Information Retrieval Systems and to find out how the performance will degrade when the irrelevant embedding is largely increased. Hence, the performance of a scaled-up version of a prototype retrieval system cannot be predicted. The recent overview of [11] does not mention *generality* as one of the required parameters for performance evaluation. However, in [10], the same authors convincingly show how the evaluation results depend on the particular content of the database. These considerations led us to re-evaluate the performance measurements for CBIR and the way these performance measures are visualized in graphs [3]. How can we make the performance measures for image queries on test databases more complete, so that the results of specific studies cannot only be used to select the better method, but can also be used to make comparisons between different system sizes and different domains? In the next section, we argue that the present measures and, in particular, the Precision-Recall Graph, are not only unsuited for comparing different systems, but are often also flawed in their use of averaging over various relevant class sizes and embedding ratios.

2 Performance Evaluation as a Quantitative Decision Support Analysis

In Information Retrieval (IR), the user having specified a query would like the system to return some or all of the items (either documents, images, or sounds) that are in some sense part of the same semantic relevant class, i.e., the relevant fraction of the database with respect to this query for this user at this time. This includes target searches [2] where one is aiming at solitary hits (relevant class size being one).

In a testing environment, the performance of the Retrieval System in its selection of database items that are retrieved should be compared to the equivalent situation where ground-truth has been constructed. An Ideal Information Retrieval System would mimic this ground-truth. Such an Ideal IR System would quickly present the user some or all of the relevant material and nothing more. The user would value this Ideal System as being either 100 percent effective or being without (0 percent) error. In this paper, we will refer to this Ideal System as the Total Recall Ideal System (TRIS). In practice, however, IR Systems are often far from ideal; generally, the query results shown to the user (a finite list of retrieved elements) are incomplete (containing only some retrieved relevant class items) and polluted (with retrieved but irrelevant items).

Let us now characterize a CBIR system using the following set of parameters:

$$number\ of\ relevant\ items\ for\ a\ particular\ query = relevant\ class\ size = c, \tag{1}$$

$$number\ of\ irrelevant\ items\ for\ a\ particular\ query = embedding\ size = e, \tag{2}$$

$$ranking\ method = m, \tag{3}$$

$$number\ of\ retrieved\ items\ from\ the\ top\ of\ the\ ranking\ list = scope = s, \tag{4}$$

$$number\ of\ visible\ relevant\ items\ within\ scope = v, \tag{5}$$

$$total\ number\ of\ items\ in\ the\ ranked\ database = database\ size = (c+e) = d. \tag{6}$$

In this set-up, the class of relevant items is considered unordered and everything that precedes a particular ranking (like user feedback) is condensed into *ranking method*. Since the relevant outcome of a particular query, v, is a function of class size, c, embedding size, e, ranking method, m, and scope, s, the performance is determined by the particular combination of the four free parameters. We will concentrate on retrieval settings where the embedding items vastly outnumber the relevant class items, $e \gg c$ and, hence, $d \approx e$:

$$v = f(c, d, m, s). \tag{7}$$

In general, the average performance will be graphed for a number of ranking methods to completely specify the retrieval system performance for a ground checked set of queries:

$$v = v_m = f(c, d, s). \tag{8}$$

In our opinion, a characterization of the Retrieval System performance should be based on the well-established decision support theory, similar to the way decision tables or contingency tables are analyzed in [6]. From a quantitative decision-support methodology, our Query By Example (QBE) situation can be characterized for each ranking method by a series of decision tables (see, for instance, [17]) or, as they are also called, contingency tables [6]. A decision

TABLE 1
Categories and Marginals for the Contingency Tables

v	$(c-v)$	c
$(s-v)$	$(d+v)-(c+s)$	e
s	$(d-s)$	d

TP	FN	P
FP	TN	N
R	NR	DB

P = Positive, N = Negative, FP = False Positive, FN = False Negative, TP = True Positive, TN = True Negative, R = Retrieved, NR = Not Retrieved, DB = Database size. In TRIS, $v = s = c$ and $TP = P = R$.

table for a ranking method represents a 2×2 matrix of a (relevant, irrelevant) versus (retrieved, not retrieved) number of items for different choices of scope, s, relevant class size, c, and embedding, e. It can also be seen as the database division according to the ground-truth versus its division according to Content-Based Information Retrieval at specific scope. The CBIR preferred choice of contingency table descriptors is given next to the Decision Support naming scheme in Table 1.

In general, the aim is to minimize a weighted combination of False Positives and False Negatives:

$$min(\alpha FP + (1-\alpha)FN) \; with \; \alpha \in [0.0, 1.0]. \tag{9}$$

2.1 Normalized Performance Measures

In this section, we will examine possible normalized views on [0, 1] or [0 percent, 100 percent] to express system performance in terms of expected success rate. In particular, we would like to show how a particular class of relevant items c is successfully retrieved within an evergrowing embedding e. The performance or relevant outcome of the query, v from (8), can be normalized by division through either c, s, or d:

$$v/c = recall = r = f(1, d/c, s/c) = f(d/c, s/c), \tag{10}$$

$$v/s = precision = p = f(c/s, d/s, 1) = f(c/s, d/s), \tag{11}$$

$$\begin{aligned} v/d = f(c/d, 1, s/d) = f(c/d, s/d) \text{ with } c/d = generality \\ = g = expected \; random \; retrieval \; rate. \end{aligned} \tag{12}$$

Equation (12) is not very useful in this form since both a low v and a high d will result into low performance figures (especially in our case of $d \approx e \gg c$).

Recall and *precision* are widely used in combination (Precision-Recall graph) to characterize retrieval performance, usually giving rise to the well-known hyperbolic graphs from high *precision*, low *recall* towards low *precision*, high *recall* values. *Precision* and *recall* values are usually averaged over precision or recall bins without regard to class size, *scope*, or embedding conditions. The fact that these are severe shortcomings can be seen from (10) and (11), where *recall* and *precision* outcomes are mutually dependent and may vary according to the embedding situation. How the dependency of *precision* and *recall* restricts the resulting outcomes is described in Section 2.3; how it affects the way p, r value pairs should be averaged is described in Section 2.4. In the next section, we will further normalize the performance description resulting in measures that are all normalized with respect to the relevant class size c and retain information about the effect of a vastly growing embedding e.

2.2 Additional Normalization of Scope

Remembering TRIS, the total recall ideal system introduced before, and observing the ratios in (10) and (11), we propose to further normalize performance figures by restricting scopes to values that have a constant ratio with respect to the class sizes involved:

$$\begin{aligned} s_r = relevant \; scope = \frac{scope}{relevant \; class \; size} \\ = \frac{s}{c} = \frac{r}{p} = \; a = constant. \end{aligned} \tag{13}$$

With this relevant scope restriction, (10) and (11) become:

$$r = f(1, d/c, ac/c) = f(1, d/c, a) = f(d/c), \tag{14}$$

$$p = r/a = f(c/ac, d/ac, 1) = f(1/a, d/ac, 1) = f(d/c). \tag{15}$$

This additional normalization of *scope*, with respect to class size c, means that the degrees of freedom for performance measures are further lowered from two to one; only *recall* or *precision* values have to be graphed versus an embedding measure. Our preferred choice for the constant a in (13) is to set $a = 1$. With this setting, one actually normalizes the whole Table 1 (now with $s = c$) by c, thus restricting one's view to what happens along the diagonal of the Precision-Recall Graph where $p = r$. This view coincides with a comparison of the retrieval system under study with TRIS (see Section 2.3).

The only remaining dependency in this set-up (apart from the method employed) is on d/c. In (12), its inverse was defined as *generality* or the expected success-rate of a random retrieval method. Although generality g is a normalized measure, we will not graph it as such because this would completely obscure the performance behavior for our case of interest, a range of $e \approx d \gg c$. Instead, we propose to graph $p = r/a$ versus $-\log_2(g)$ or $\log_2(d/c)$ to make the generality axis unbounded by giving equal space to each successive doubling of the embedding with respect to the relevant class size.

2.2.1 PR Graphs: The Restriction to a Generality Plane and Addition of Generality Information

The general three-dimensional retrieval performance characterization can be presented in 2D as a set of Precision-Recall graphs (for instance, at integer logarithmic generality levels) to show how the p, r values decline due to successive halving of the relevant fraction. In this paper, another attractive plane in three-dimensional Generality-Precision-Recall space, the Precision=Recall plane (see Fig. 4), will be advocated for the characterization of system performance.

In general, Precision-Recall graphs have been used as if the generality level would not matter and any p, r, g curve can be projected on a $g = constant$ plane of the three-dimensional performance space. However, our experiments reported in [4] show (at least for Narrow-Domain CBIR embeddings) that it does matter and, therefore, Precision-Recall graphs should only be used to present performance evaluations when there is a more or less constant and clearly specified generality level. Only the Total Recall Ideal

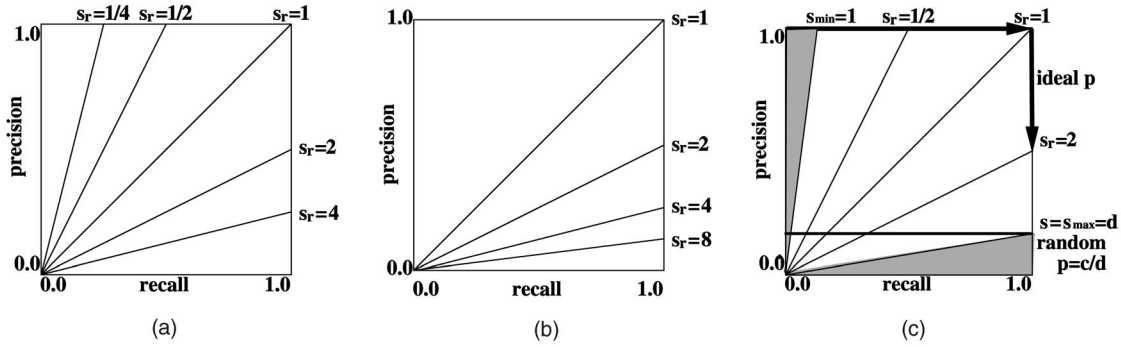

Fig. 2. (a) Lines along which p, r values are located at relevant class size = 4 for several scopes. (b) Lines along which p, r values for retrieved relevant class size = 1 (relevant class of 2, 1 used for query, max 1 for retrieval) are located. (c) p, r values for ideal retrieval are 1, r for $r < 1$; for scope size $>$ relevant class size, p drops slowly toward the random level, c/d.

System (TRIS) as described for the PR graph is insensitive to generality by definition.

2.3 Scope Graphs Contained in P-R Graphs: Normalized Scope

In this section, we will show that although many authors present Precision-Scope or Recall-Scope graphs, these graphs can be directly extracted from the Precision-Recall graph.

Information about the effect of changing the *scope* on the measured *precision* and *recall* values can be made visible in the Precision-Recall graph by taking into account that possible *precision*, *recall* outcomes are restricted to lay on a line in the PR-graph radiating from the origin. This is due to the fact that the definitions of the system parameters *precision* (see (11)) and *recall* (see (10)) have the same numerator v and are, therefore, not independent. The dependent pair of p, r values and its relation to *scope* becomes even more pronounced when *scope* is normalized with respect to the number of relevant items, as defined by (13). We call this measure *relevant scope* and present p, r values accompanied by their relevant scope line (radiating from the origin). So, for each scope $s = a \cdot c$ with an arbitrary positive number a, $s_r = a$, and the p, r values are restricted to the line $p = r/a$. In Fig. 2a, several constant scope lines for retrieval of a relevant class of four additional relevant class members are shown.

With these relevant scope lines drawn in the Precision-Recall graph, one understands much better what the p, r values mean. In the ideal case (see Fig. 2c), *precision* p will run along $p = 1.0$ for *recall* $r \in [0.0, 1.0)$ and reach $p, r = 1.0, 1.0$ (the TRIS point) when *scope* equals relevant class size ($s = c$); for scopes greater than relevant class size, *precision* will slowly drop from $p = 1.0$ along $r = 1.0$ until the random level $p = c/d$ at $s = d$ is reached.

Also depending on relevant class size, the region to the left of $p = r/c$ cannot be reached (solving the difficulty in PR-graphs for selecting a *precision* value for $recall = 0.0$) as well as the region below $p = dr/c$. This means that for the smallest relevant class of two members where one of the relevant class members is used to locate its single partner, the complete upper-left half of the PR graph is out of reach (see Fig. 2b).

Because the diagonal $s = c$ line presents the hardest case for a retrieval system (last chance of *precision* being max 1.0 and first chance of *recall* being max 1.0) and is the only line that covers all relevant class sizes (see Fig. 2b), the best total recall system performance presentation would be the $p = r$ plane in the three-dimensional GReP Graph (Generality-Precision-Recall Graph).

2.4 Radial Averaging of Precision, Recall Values

For system performance, one normally averages the discrete sets of *precision* and *recall* values from single queries by averaging *precision*, *recall* values without paying attention to the *generality* or *scope* values associated with those measurements. To compensate for the effect generality values have on the outcome of the averaging procedures, different ways of averaging are applied, like the micro and macroaveraging used in text-retrieval [16]. In the critical review [13], the authors state, with respect to averaging *precision* and *recall* values within the same database, that *precision* values should be averaged by using constant *scope* or cut-off values, rather than using constant *recall* values.

The fact stressed in Section 2.3 that p, r results have associated *generality* and relevant scope values also has implications for the way average PR curves should be made up. Instead of averaging p, r values within recall or scope bins, one should average p, r values along constant relevant scope lines and only those that share a common *generality* value. When averaging for query results obtained from a constant size test database, the restriction to averaging over outcomes of queries with constant relevant class sizes (constant generality value) will automatically result in identical micro and macroaverages. The view expressed by [13] should, therefore, even be refined: the recipe of averaging measured *precision*, *recall* values over their associated constant *scope* values only should further be refined to our recipe of averaging p, r values over constant associated s_r, g values only.

An example of the way we determine an average p, r curve out of two individual curves with a shared generality value is given in Fig. 3. The figure illustrates how averaging *recall* values in constant precision boxes (pbox-averaging) overestimates *precision* at low recall values while underestimating it at high recall values, whereas averaging of *precision* values in constant recall boxes (rbox-averaging) underestimates *precision* at low recall while overestimating it at high recall values. In case of averaging discrete *precision*, *recall* values, the errors introduced by not averaging radially (along constant relevant scope s_r) can be even more dramatic.

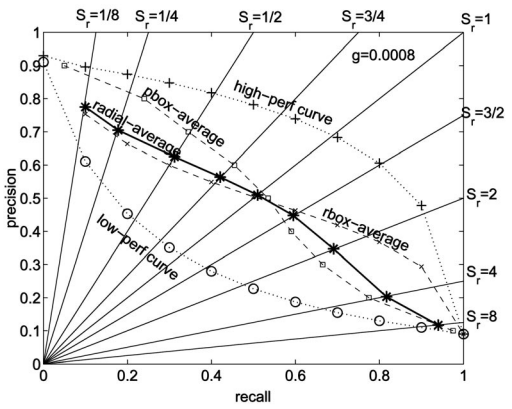

Fig. 3. Average PR-curves obtained from a low and a high-performing PR-curve for two queries with class size 16 embedded in 21.094 images. The figure shows how large the difference can be between radial averaging compared to either precision-box averaging or recall-box averaging.

3 A Unified View on Total Recall Performance

In the light of our user's Total Recall Ideal System, introduced earlier, one can highlight the system performance by restricting to the diagonal plane in Generality-Recall-Precision space. This plane contains the *precision*, *recall* values where relevant scopes are one ($s = c$). Therefore, in this case, we obtain the most unifying view on system performance with $recall = precision$.

The two-dimensional graph, showing p, r values as a function of g (on a logarithmic scale), will be called the Generality-Recall=Precision Graph, GRiP Graph for short (see Fig. 4a and Fig. 4b). For Total Recall studies, one could present several GRiP related graphs for planes in the GReP Graph, where $recall = n{\cdot}precision$ corresponding to the situation where the scope for retrieval is a multiple of the relevant class size ($s_r = n$). We shall denote these Generality-Recall=nPrecision Graphs as GRnP Graphs; obviously, the GRiP Graph corresponds to the GR1P Graph. By showing system performance for GR1P and GR2P, indicating the performance for $s_r = 1$ and $s_r = 2$, the usability of the system for Total Recall would be well characterized. Its function can be compared to the Bull's Eye Performance (BEP) measure used in MPEG-7 for shape and motion descriptors [9], but extended to include the effect of *generality* on the performance. Another well-known associated overall measure (but without taking generality into account) is van Rijsbergen's E-measure, which we discuss in the next section.

3.1 A Generalization of Van Rijsbergen's E-measure

To show how our GRiP Graph fits into the Information Retrieval literature, we will discuss van Rijsbergen's E-measure that, for a specific parameter setting, will be shown to be equivalent to a GRiP value for a specific generality value.

The parameterized E-measure of van Rijsbergen [18]:

$$E = 1 - \frac{1}{\alpha(1/p) + (1-\alpha)(1/r)} \tag{16}$$

is a normalized Error-measure where a low value of α favors *recall* and a high value of α favors *precision*. E will be 0 for an ideal system with both *precision* and *recall* values at 1 (and, in this case, irrespective of α). Van Rijsbergen [18] favors the setting of $\alpha = 0.5$, a choice giving equal weight to *precision* and *recall* and giving rise to the normalized symmetric difference as a good single number indicator of system performance (or rather system error):

$$Error = E(\alpha = 0.5) = 1 - \frac{1}{(1/2p) + (1/2r)}. \tag{17}$$

The problem with this E-measure is fourfold:

- An intuitive best value of 1 (or 100 percent) is to be preferred; this can easily be remedied by inverting the [1,0] range by setting E to its range inverted and more simple form:

$$Effectiveness = 1 - E(\alpha = 0.5) = \frac{1}{(1/2p) + (1/2r)}. \tag{18}$$

- An indication of generality (database fraction of relevant class size) is missing completely.

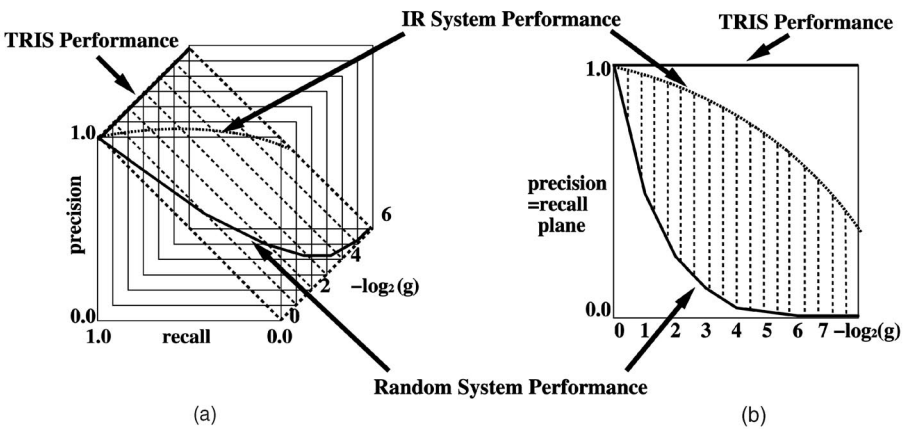

Fig. 4. (a) The 3D GReP Graph with the $p = r$ plane (with random and $s = c$ results for different generality values) holding the GRiP Graph. (b) The 2D GRiP Graph $p = r$. Values for scope size = relevant class size as a logarithmic function of generality.

- An indication of expected random retrieval performance is missing.
- An indication of expected scope is missing; for the user, the length of the list to be inspected is very important, so knowing *precision* as a function of *scope* is highly appreciated.

To best compare Information Retrieval System Performance between competing methods and among systems, one would rather like to use a normalized triple like $\{-\log_2(g), r, p\}$ (log generality, recall, and precision). Then, (18) using (10) and (11) becomes:

$$Effectiveness = \frac{2v}{s+c}. \tag{19}$$

In the case of the $p = r$ plane, where $s = c$ and $p = r = v/c$, this becomes:

$$Effectiveness = \frac{2v}{c+c} = \frac{v}{c} = p = r. \tag{20}$$

This shows that van Rijsbergen's inverted E-measure at $\alpha = 0.5$ will turn into our GRiP Graph (Generality-Recall= Precison Graph) by taking into account that it is a function of *generality*:

$$Effectiveness(g) = \frac{v(g)}{c} = p(g) = r(g). \tag{21}$$

The system performance would then be given by $E(g)$ or, rather, $E(-\log_2 g)$ (*effectiveness* as a function of generality).

By subtracting random performance or generality g (12) from the E-measure (21), a simple performance measure E^* is obtained that indicates the gain with respect to a random retrieval method; for large generalities, in most cases, $E^* \approx E$ when $p(g) \gg g$ and $g \approx 0.0$. For comparison with TRIS at $s = c$, $p(g) = r(g)$ and

$$E^*(g) = r(g) - g = p(g) - g, \tag{22}$$

which corresponds to the shaded area in Fig. 4b. $E^*(g)$ penalizes the use of small embeddings in retrieval tests, but will approximate $E(g)$ for large embeddings.

The fact that one of the CBIR performance descriptors, *generality*, is graphed logarithmically prevents us from characterizing retrieval performance by a single overall measure by taking the area under the graph, like the A-measure suggested by Swets as early as 1963 [15], for Information Retrieval and which was shown to correspond to the area under the Recall-Fallout Graph by Robertson in [14].

4 Laboratory Systems versus Practical Systems

We have shown that, for a complete performance evaluation, one has to carry out controlled retrieval tests with queries for which ground-truth can provide the relevant class sizes. The performance is measured for various ranking methods within a range of *scope* and *generality* values.

Since it is often too costly and labor intensive to construct the complete ground-truth for the queries used, we will indicate what could be done in terms of evaluation when knowledge about relevant class sizes c and, as a result, *recall* and *generality* values are missing.

First, let us make a distinction between Laboratory and Practical CBIR systems. We propose to reserve the name Laboratory CBIR system for those performance studies where complete ground-truth has become available. For these systems, a complete performance evaluation in the form of Generality-Recall-Precision Graphs for a set of test queries and for a number of competing ranking methods can be obtained.

Any CBIR retrieval study that lacks complete ground-truth will be called a Practical system study. In Practical system evaluation, one normally has a set of queries and a database of known size d. Because ground-truth is missing, relevant class size c is unknown. The only two free controls of the experimenters are the scope s and the ranking method m. Relevance judgments have to be given within the scopes used to determine the number of relevant answers. Of the three Laboratory system evaluation parameters, *precision*, *recall*, and *generality*, only $precision = v/s$ is accurately known. For *recall*, due to knowing v but not c, only a lower bound $v/(d - s + v)$ is known. For *generality*, only a lower bound $g = v/d$ is known. In general, for practical studies, one characterizes the performance as Precision-Scope Graphs or one uses single measures obtained from the weighted ranks of the relevant items within scope.

The problem with any Practical system study is that one cannot interpret the results in terms of "expected completeness" (recall) and the results are therefore only useful in terms of economic value of the system: how many items will I have to inspect extra, to obtain an extra relevant item? Actually, with some extra effort, the analysis of a Practical system can be enhanced to that of an estimated Laboratory system by using the fact that *generality*, in terms of relevant fraction, is identical to the expected *precision* (see (12)) when using a random ranking method.

Experimenters that have access to the ranking mechanism of a retrieval system can thus obtain estimates for generality g and, hence, estimates for relevant class size c and recall r to complete their performance evaluation. The extra effort required would be the making of relevance judgments for a series of randomly ranked items within some long enough scope for each query.

5 Conclusions

We surveyed how the role of embeddings in Content-Based Image Retrieval performance graphs is taken care of and found it to be lacking. This can be overcome by adding a generality component. We also noted that one is not aware of the scope information present in a Precision-Recall Graph and the lack of comparison with random performance. The present practice of averaging *precision*, *recall* values in recall or precision boxes conflicts with the way *precision* and *recall* are dependently defined.

We conclude that Precision-Recall Graphs can only be used when plotting *precision*, *recall* values obtained under a common, mentioned, *generality* value which coincides with the random performance level. Therefore, to complete performance space we extended the traditional 2D Precision-Recall graph to the 3D GReP Graph (Generality-Recall-Precision Graph) by adding a logarithmic generality dimension. Moreover, due to the dependency of *precision* and *recall*, their combined values can only lay on a line in the PR Graph determined by the *scope* used to obtain their values. Scopes, therefore, can be shown in the PR Graph as a set of

radiating lines. A normalized view on scope, relevant scope, makes the intuitive notion of scope much simpler. Also, averaging *precision*, *recall* values should be done along constant relevant scope lines and only for those p, r values that have the same *generality* value.

Our recipe is as follows: From the ranking lists d, c, s, and v are determined for each query and from these an associated set of normalized values p, r, g, s_r are computed. p, r, g values can be used to construct the GReP graph. Note that the s_r value is only needed in the case where one wants to average over sets of p, r values. Averaging should be restricted to those p, r measurements that share common s_r, g values.

For Total Recall System performance, we advocated a comparison with the Total Recall Ideal System performance by using a special plane in the GReP Graph, the 2D GRiP Graph (Generality-Recall=Precision Graph), showing the diagonal of the PR Graphs (*recall* and *precision* being equal on the diagonal of the PR graph) as a logarithmic function of generality. In this way, statements can be made about what to expect from the system for the retrieval of specific relevant class sizes in a range of embedding database sizes. We also make a distinction between Laboratory CBIR Systems and Practical CBIR Systems: for Laboratory Systems, complete ground-truth is available, for Practical Systems, it is lacking but can be estimated.

The extensions to performance graphs, suggested in this paper, make it possible to better compare performance figures within growing CBIR systems (by explicit mentioning of the generality level) and make it possible to infer *precision* and *recall* as a function of normalized scope, reducing the need for additional *precision* or *recall* versus *scope* graphs next to Precision-Recall graphs. For the evaluation of the performance, in relation to continually growing database sizes, the GRiP Graph (Generality-Recall=Precision Graph) offers the best overall IR System performance overview since this graph shows how well the system in question approaches TRIS. A simple variant and generalization of van Rijsbergen's E-measure was shown to describe retrieval effectiveness in the same way.

Finally, a critical note on the usefulness of generality and relevant scope performance measures: Because the degradation effect of a growing embedding depends on both its size and its nature, a real comparison between systems, to determine the better ranking methods, is only possible in publicly available benchmarks with ground-truth test queries, like Benchathlon [7]. We have therefore offered our test database of B/W portraits and logo's plus ground-truth and test queries for inclusion in such a benchmark.

REFERENCES

[1] C. Baumgarten, "A Probabilistic Solution to the Selection and Fusion Problem in Distributed Information Retrieval," *Proc. Int'l ACM Special Interest Group on Information Retrieval Conf. (SIGIR '99),* pp. 46-253, 1999.

[2] I.J. Cox, M.L. Miller, T. Minka, T.V. Papathomas, and P.N. Yianilos, "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments," *IEEE Trans. Image Process,* vol. 9, no. 1, pp. 20-37, 2000.

[3] D.P. Huijsmans and N. Sebe, "Extended Performance Graphs for Cluster Retrieval," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '01),* pp. 26-31, 2001.

[4] D.P. Huijsmans and N. Sebe, "Content-Based Indexing Performance: A Class Size Normalized Precision, Recall, Generality Evaluation," *Proc. IEEE Int'l Conf. Image Processing (ICIP '03),* pp. 733-736, 2003.

[5] *The SMART Retrieval System,* G. Salton, ed. Prentice Hall, 1971.

[6] D.V. Gokhale and S. Kullback, *The Information in Contingency Tables.* New York: Marcel Dekker, Inc., 1978.

[7] http://www.benchathlon.net.

[8] D.P. Huijsmans, N. Sebe, and M.S. Lew, "A Ground-Truth Training Set for Hierarchical Clustering in Content-Based Image Retrieval," *Proc. Fourth Int'l Conf. Advances in Visual Information Systems,* pp. 500-510, 2000.

[9] *IEEE Trans. Circuits and Systems for Video Technology,* special issue: MPEG-7, vol. 11, no. 6, 2001.

[10] H. Müller, S. Marchand-Maillet, and T. Pun, "The Truth about Corel—Evaluation in Image Retrieval," *Proc. Int'l Conf. Image and Video Retrieval (CIVR '02),* pp. 38-49, 2002.

[11] H. Müller, W. Müller, D.M. Squire, S. Marchand-Maillet, and T. Pun, "Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals," *Pattern Recognition Letters,* vol. 22, pp. 593-601, 2001.

[12] K. Porkaew, K. Chakrabarti, and S. Mehrotra, "Query Refinement for Multimedia Similarity Retrieval in MARS," *ACM Multimedia,* pp. 235-238, 1999.

[13] V. Raghavan, P. Bollmann, and G.S. Jung, "A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance," *ACM Trans. Information Systems,* vol. 7, pp. 205-229, 1989.

[14] S.E. Robertson, "The Parametric Description of Retrieval Tests. Part II: Overall Measures," *J. Documentation,* vol. 25, no. 2, pp. 93-107, 1969.

[15] J.A. Swets, "Information Retrieval Systems," *Science,* vol. 141, pp. 245-250, 1963.

[16] J. Tague-Sutcliffe, "The Pragmatics of Information Retrieval Experimentation, Revisited," *Information Processing and Management,* vol. 28, no. 4, pp. 467-490, 1992.

[17] *Handbook of Medical Informatics,* J.H. van Bemmel and M.A. Musen, eds. Springer, 1997.

[18] C.J. van Rijsbergen, *Information Retrieval.* London: Butterworths, 1979.

[19] N. Vasconcelos and A. Lippman, "A Probabilistic Architecture for Content-Based Image Retrieval," *Proc. CVPR '00,* pp. 216-221, 2000.

[20] *Proc. Text REtrieval Conf.,* E.M. Voorhees and D. Harman, eds., 1999.

Dionysius P. Huijsmans received the PhD degree in mathematics and physics from the University of Amsterdam in 1982. From 1982 to 1985, he did postdoctoral research aimed at three-dimensional reconstruction from seril sections at the Laboratory for Medical Physics in Amsterdam. Since 1985, he has been with Leiden University in the Netherlands as an assistant professor at the LIACS and ASCI Research school. His main research area is computer imagery and multimedia searches with noisy keys.

Nicu Sebe is an assistant professor in the Faculty of Science, University of Amsterdam, The Netherlands, where he is doing research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He is the author of the book *Robust Computer Vision—Theory and Applications* (Kluwer, April 2003) and of the upcoming book *Machine Learning in Computer Vision*. He was a guest editor of a CVIU special issue on video retrieval and summarization (December 2003) and was the cochair of the fifth ACM Multimedia Information Retrieval Workshop, MIR '03 (in conjunction with ACM Multimedia 2003), and of the first Human Computer Interaction Workshop, HCI '04 (in conjunction with ECCV 2004). He acts as the cochair of the sixth ACM Multimedia Information Retrieval Workshop, MIR '04 (in conjunction with ACM Multimedia 2004). He was also the technical program chair for the International Conference on Image and Video Retrieval, CIVR 2003. He has published more than 50 technical papers in the areas of computer vision, content-based retrieval, pattern recognition, and human-computer interaction and has served on the program committee of several conferences in these areas. He is a member of the IEEE and the ACM.