

Complete Performance Graphs in Probabilistic Information Retrieval

N. Sebe¹, D.P. Huijsmans², Q. Tian³, and T. Gevers¹

¹Faculty of Science, University of Amsterdam, The Netherlands

²Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

³University of Texas at San Antonio, San Antonio, USA

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 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. We also propose new performance graphs for total recall studies in a range of embeddings.

1 Introduction

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 [1] and van Rijsbergen [2] as well as the proceedings of the annual ACM SIGIR and NIST TREC conferences.

In probabilistic text retrieval [2], TREC [3], and MPEG-7 descriptor performance evaluation [4], authors often go for single measure performance characterizations. These single measures are based on ranking results within a limited scope and in most cases they do take 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.

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 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 [5–7]. 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 IR 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 [8] does not mention *generality* as one of the required parameters for performance evaluation. However, in [9] the 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 [10]. How can we make the performance measures for image queries on test databases more complete, so that 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?

2 Performance Evaluation Elements

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% effective or being without (0%) error. In [11], we referred 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).

We characterize a CBIR system using the following set of parameters:

$$\text{number of relevant items for a particular query} = \text{relevant class size} = c \quad (1)$$

$$\text{number of irrelevant items for a particular query} = \text{embedding size} = e \quad (2)$$

$$\text{ranking method} = m \quad (3)$$

$$\text{number of retrieved items from the top of the ranking list} = \text{scope} = s \quad (4)$$

$$\text{number of visible relevant items within scope} = v \quad (5)$$

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

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 the *ranking method*. Performance is determined by the particular combination of the 4 free parameters, 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 . However, 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. We also concentrate on retrieval settings where the embedding items vastly outnumber the relevant class items, $e \gg c$ and hence $d \approx e$:

$$v = v_m = f(c, d, s). \quad (7)$$

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 [12]. 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 [13] or, as they are also called, contingency tables [12]. A decision table for a ranking method represents a 2×2 matrix of (*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.

v	$(c - v)$	c	TP	FN	P
$(s - v)$	$(d + v) - (c + s)$	e	FP	TN	N
s	$(d - s)$	d	R	NR	DB

Table 1. Categories and marginals for the contingency tables: 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 $TP = P = R$.

The performance or relevant outcome of the query, v from Eq. (7), can be normalized by division through either c , s , or d :

$$v/c = recall = r = f(1, d/c, s/c) = f(d/c, s/c) \quad (8)$$

$$v/s = precision = p = f(c/s, d/s, 1) = f(c/s, d/s) \quad (9)$$

$$v/d = f(c/d, 1, s/d) = f(c/d, s/d) \quad (10)$$

with $c/d = generality = g = expected\ random\ retrieval\ rate$.

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. That these are severe shortcomings can be seen from (8) and (9) where *recall* and *precision* outcomes are mutually dependent and may vary according to the embedding situation. To address these shortcomings, we propose to further normalize performance figures by restricting scopes to values that have a constant ratio with respect to the class sizes involved:

$$s_r = relevant\ scope = \frac{scope}{relevant\ class\ size} = \frac{s}{c} = \frac{r}{p} = a = constant \quad (11)$$

With this relevant scope restriction, Eqs. (8) and (9) become:

$$r = f(1, d/c, ac/c) = f(1, d/c, a) = f(d/c) \quad (12)$$

$$p = r/a = f(c/ac, d/ac, 1) = f(1/a, d/ac, 1) = f(d/c). \quad (13)$$

This additional normalization of *scope* with respect to class size c means that the degrees of freedom for performance measures are further lowered from 2 to 1; only *recall* or *precision* values have to be graphed versus an embedding measure. Our preferred choice for the constant a in Eq. (11) is to set $a = 1$. With this

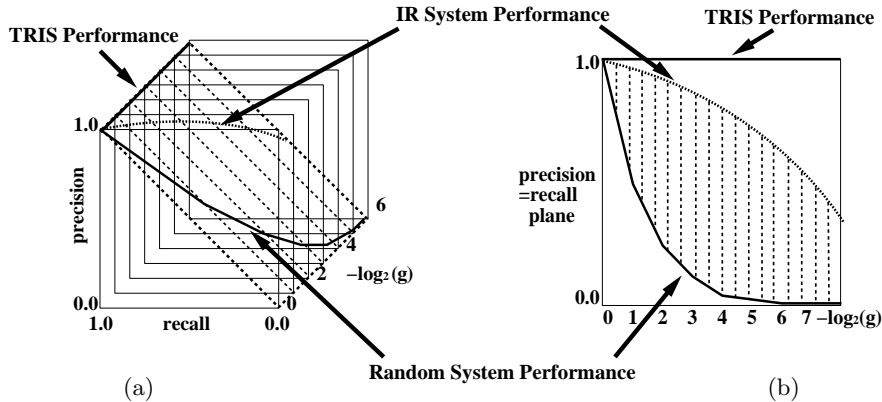


Fig. 1. (a) The 3D GReP Graph with the $p = r$ plane (with random and $s = c$ results for different generality values) holding the GRiP Graph; (b) The 2D GRiP Graph: $p = r$ values for scope size=relevant class size as a logarithmic function of generality.

setting one actually normalizes the whole Table 1 (now with $s = c$) by c , thus restricting ones view to what happens along the diagonal of the Precision-Recall Graph where $p = r$.

The only remaining dependency in this set-up (apart from the method employed) is on d/c . In Eq. (10) 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 $-\log_2(g)$ to make the generality axis unbounded by giving equal space to each successive doubling of the embedding with respect to the relevant class size. We obtain thus the 3D Generality-Recall-Precision (GReP) graph (see Figure 1(a)).

The general 3D 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. 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 Figure 1(b)). 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= n Precision Graphs as GRnP Graphs; obviously the GRiP Graph corresponds to the GR1P Graph.

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

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

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 *precision* (Eq. (9)) and *recall* (Eq. (8)) 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 Eq. (11). Therefore, we present p, r values accompanied by their relevant scope line (radiating from the origin). So for each scope $s = a \cdot c$ with a an arbitrary positive number, $s_r = a$ and the p, r values are restricted to the line $p = r/a$. In Figure 2(a) we show several constant scope lines for retrieval of a relevant class of four additional relevant class members.

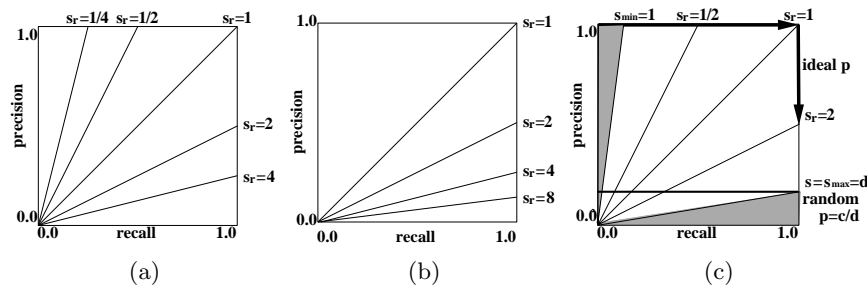


Fig. 2. (a) Lines along which p, r values are located at relevant class size=4 for several scopes; (b) Lines along which p, r values for retrieved relevant class size=1 (relevant class of 2, 1 used for query, max 1 for retrieval) are located; (c) p, r values for ideal retrieval are $1, r$ for $r < 1$;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

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 Figure 2(c)), *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 2 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 Figure 2(b)).

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 Figure 2(b)), the best total recall system performance presentation would be the $p = r$ plane in the three-dimensional GReP Graph (Generality-Precision-Recall Graph).

2.2 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 macro-averaging used in text-retrieval [15]. In the critical review [16], 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 2.1, 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 macro-averages. The view expressed by [16] 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 2 individual curves with a shared generality value is given in Figure 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.

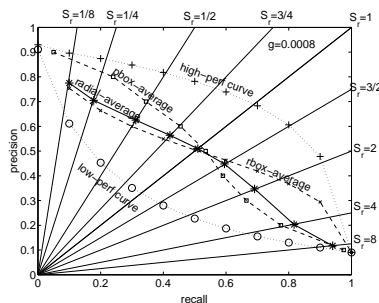


Fig. 3. Average PR-curves obtained from a low- and a high-performing PR-curve for 2 queries with class size 16 embedded in 21.094 images: the figure shows how large the difference can be between radial averaging compared to either precision-box (pbox) averaging or recall-box (rbox) averaging.

3 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 Eq. (10)) 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.

4 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.

References

1. Salton, G.: The SMART retrieval system. Prentice Hall (1971)
2. van Rijsbergen, C.: Information Retrieval. Butterworths, London (1979)
3. Voorhees, E.M., Harman, D.: Proc. TREC. (1999)
4. MPEG-7. Special Issue of IEEE Trans. Circuits and Systems for Video Technology **11** (2001)
5. Porkaew, K., Chakrabarti, K., Mehrotra, S.: Query refinement for multimedia similarity retrieval in MARS. In: ACM Multimedia. (1999) 235–238
6. Vasconcelos, N., Lippman, A.: A probabilistic architecture for content-based image retrieval. In: CVPR. (2000) 216–221
7. Baumgarten, C.: A probabilistic solution to the selection and fusion problem in distributed information retrieval. In: SIGIR. (1999) 246–253
8. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Performance evaluation in content-base image retrieval: Overview and proposals. Pattern Recog. Letters **22** (2001) 593–601
9. Müller, H., Marchand-Maillet, S., Pun, T.: The truth about Corel - Evaluation in image retrieval. In: CIVR 2002. (2002) 38–49
10. Huijsmans, D., Sebe, N.: Extended performance graphs for cluster retrieval. In: CVPR. (2001) 26–31
11. Huijsman, D.P., Sebe, N.: How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. IEEE Trans. on PAMI, to appear (2004)
12. Gokhale, D., Kullback, S.: The Information in Contingency Tables. M. Dekker (1978)
13. van Bemmelen, J.H., Musen, M.A.: Handbook of Medical Informatics. Springer (1997)
14. Huijsmans, D., Sebe, N.: Content-based indexing performance: A class size normalized precision, recall, generality evaluation. In: ICIP. (2003) 733–736
15. Tague-Sutcliffe, J.: The pragmatics of information retrieval experimentation, revisited. Information Processing and Management **28** (1992) 467–490
16. Raghavan, V., Bollmann, P., Jung, G.: A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Information Systems **7** (1989) 205–229