Noname manuscript No.
(will be inserted by the editor)

MEAN NORMALIZED RETRIEVAL ORDER (MNRO). A NEW CONTENT-BASED IMAGE RETRIEVAL PERFORMANCE MEASURE

Savvas A. Chatzichristofis · Chryssanthi Iakovidou · Yiannis S. Boutalis · Elli Angelopoulou

Received: date / Accepted: date

Abstract The results of a content based image retrieval system can be evaluated by several performance measures, each one employing different evaluation criteria. Many of the methods used in the field of information retrieval have been adopted for use in image retrieval systems. This paper reviews the most widely used performance measures for retrieval evaluation with particular emphasis on the assumptions made during their design. More specifically, it focuses on the design principles of the commonly used Mean Average Precision (MAP) and Average Normalized Modified Retrieval Rank (ANMRR), pinpointing their limitations. It also proposes a new performance measure for image retrieval systems, the *Mean Normalized Retrieval Order (MNRO)*, whose effectiveness is demonstrated through a wide range of experiments. Initial experiments were conducted on artificially produced query trials and evaluations. Experiments on a large database demonstrate the ability of MNRO to take into account the generality of the queries during the retrieval procedure. Furthermore, the results of a case study show that the proposed performance measure is closer to human evaluations, in comparison to MAP and ANMRR. Lastly, in order to encourage researchers and practitioners to use the pro-

Savvas A. Chatzichristofis
Department of Electrical & Computer Engineering, Democritus University of Thrace, Xanthi, Greece E-mail: schatzic@ee.duth.gr

Chryssanthi Iakovidou
Department of Electrical & Computer Engineering, Democritus University of Thrace, Xanthi, Greece E-mail: ciakovid@ee.duth.gr

Yiannis S. Boutalis
Department of Electrical & Computer Engineering, Democritus University of Thrace, Xanthi, Greece, Also Visiting scholar at the Department of Electrical, Electronic and Communication Engineering, Chair of Automatic Control, Friedrich-Alexander University of Erlangen-Nuremberg, 91058 Erlangen, Germany. E-mail: ybout@ee.duth.gr

Elli Angelopoulou
Department of Computer Science, Pattern Recognition Lab, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany E-mail: elli@immd5.informatik.uni-erlangen.de

posed performance measure, we present the experimental results produced by a large number of state of the art descriptors applied on three well-known benchmarking databases.

Keywords Image Retrieval Performance Measures, Mean Average Precision, Average Normalized Modified Retrieval Rank

1 INTRODUCTION

The objective of an image retrieval system is to retrieve images in rank order, where the rank of an image is determined by its relevance to the query at hand [1]. The image retrieval process can be executed either with the use of a *keyword* 'upon' the images (Keyword Based Image Retrieval) or with the use of low-level characteristics exported from the image's visual content (Content Based Image Retrieval). Content based image retrieval (CBIR) is defined as any technology that, in principle, helps to organize digital image archives by their visual content. According to this definition, anything ranging from an image similarity function to a robust image annotation engine, falls under the purview of CBIR [2].

The performance of an information retrieval system, in general, is typically measured by using either user-centered evaluation methods or system-oriented evaluation frameworks. User-centered evaluation is an interactive method. The users judge the success of a query directly after the query. This includes more than just technical aspects, since a large number of factors influence the user's judgment on the entire retrieval system [3]. Many investigators have highlighted the advantages offered by user-centred evaluation methods in image, music-audio and text retrieval [4][5]. However, user-centered evaluations can be subjective, given that different users might judge the same retrieval result in quite distinct ways. Even the same user might judge the same result differently at different times [6]. Another drawback of user-centered evaluation is that it is very hard to get a large number of user comparisons as their collection is quite time consuming [7].

Thus, CBIR systems as well as music-audio retrieval systems have focused on a system-oriented evaluation framework. Image retrieval systems are primarily evaluated against a known ground truth dataset. A benchmark image database is used in these evaluations. Most of the relevance sets for system-oriented evaluation are based on real user judgments and are thus also subjective reflecting the opinion of one user at a particular time. Classic examples of such databases are the Wang [8] database, the UCID database [9], the Nister database [10] as well as the MIRFlicker database [11]. Each database is comprised of a number of N images and Q queries. Queries are images used as input to the retrieval system in order to evaluate its performance. For each query a number of images with visual similarity which are considered as the ground truth is given.

One can classify information retrieval systems into two categories, Boolean and item-ranking, based on the employed retrieval method. Boolean type retrieval systems, also known as classification systems, return only a set of items that are similar to the query items. A classification system can be completely described with four numbers: the size of the database, the total number of the retrieved images, the total size of the relevance set and the number of relevant image retrieved.

Image retrieval systems, on the other hand, return rankings and not sets, so we need performance measures over rankings. A system's performance is calculated using a technique that evaluates the rank of the images which form the ground truth for all the queries. Many of the performance measures that are used in the field of information retrieval have been adopted in order to evaluate image retrieval results. Section 2 presents an overview of the most common system-oriented performance measures for evaluating retrieval systems. Among these measures, the Mean Average Precision (MAP) is the most frequently used one. Still, the Averaged Normalized Modified Retrieval Rank (ANMRR) [12], which is based on MPEG-7 [13] [14], alongside with a set of other descriptors, is considered the most suitable for image retrieval systems.

However, as it is shown in this paper, in developing these two performance measures, various assumptions were made which created drawbacks with respect to the evaluation of image retrieval systems. CBIR alone is very unlikely to fulfill the user needs in searching image archives. Although, due to recent achievements in object detection and recognition, semantic analysis and understanding of images is much further developed, the desired retrieval requirements are not satisfiable [15].

CBIR systems typically extract several low level features from the images, mapping the visual content into a new space called the feature space. Features for a given image are stored in a descriptor that can be used for retrieving similar images. The key to a successful retrieval system is to choose the right features that represent the images as accurately as possible. The main problem arises from the fact that these low level features are neither rich enough, nor discriminative enough for describing the objects present in an image . Thus, CBIR is notoriously noisy, especially when global undiscriminative low-level features are employed. For example, a query image of a red tomato on a white background would retrieve a red pie-chart on white paper. If the query image happens to have a low generality, especially in large databases, early rank positions may be dominated by spurious results such as the pie-chart, which may even be ranked before tomato images [16]. Even if the retrieval approach adopts richer low-level features, such as visual words, the low discriminative power of the images themselves may affect the quality of the results [17]. Hence, it is quite common in CBIR systems that images having similar visual content but distinct semantic meaning to the query image to appear often among the early retrieval positions. This is a problem that is very particular and common in image retrieval and, rather rare in text retrieval (for example in case of synonyms). For this reason, the performance measures of CBIR systems should not be so biased at the top-10 or top-20 positions. Rather, a better technique is to use a threshold which is directly connected to either the generality of the query, or the number of items relevant to the query.

Another distinguishing characteristic between CBIR and information retrieval is the manner in which these two systems display their results. CBIR methods typically rank the whole collection via a distance measure and show the results as a table of images on the screen (see for example Google Images or Microsoft Bing Images) instead of in a list as in text results. People have the ability to recognize the relevance of a photographic result at a single glance, something that is not easily feasible in text retrieval. Thus, in CBIR small differences in the ranks should not be punished as strictly as in text retrieval.

MAP shows a tendency to be consistently correlated in the first 10 to 20 results. On the other hand, ANMRR, which was proposed for use predominantly in image retrieval systems, recognizes the specificity of the CBIR system's results and gives a bias to the recall at K, where K is directly correlated to the size of the ground truth of the query. A possible drawback of the ANMRR performance measure relies on the fact that if the image appears after the K^{th} position it is considered as not having been retrieved. This principle of operation of ANMRR does not allow for a comprehensive evaluation of recall-oriented tasks.

Another disadvantage of both MAP and ANMRR it that they do not take into account the size of the image database. For the same ground truth, the system performance degrades for larger image databases. Thus, the behavior of a scaled-up version of the system cannot be predicted. A detailed description of these 2 performance measures, an outline of the assumptions made during their design, as well as a description of the drawbacks caused by these assumptions is given in Section 3. A preliminary version of this work has been presented in [18].

To alleviate some of the limitations of MAP and ANMRR, we propose a new image retrieval performance measure which is described in details in Section 4. The new performance measure, which is called **M**ean **N**ormalized **R**etrieval **O**rder (MNRO), is rating each result with a value in the range $[0, 1]$ and does not carry the drawbacks of the previous performance measures. The effectiveness of MNRO is examined on artificial query trials, on a considerably large database and on three benchmark databases. These experiments demonstrate the ability of the proposed performance measure to take into account the generality of the queries during the retrieval procedure. MNRO's capability to mimic human evaluations of retrieval systems is also evaluated. In a case study involving 30 individuals, it is shown that the proposed performance measure is closer to the human's evaluations, in comparison to MAP and ANMRR. The experimental evaluation is described in details in Section 5.

Finally, the conclusions are drawn in Section 6. The proposed performance measure has been implemented and used in evaluating the retrieval results of the img(Rummager) system [19], which can be found on-line[1].

[1] http://www.img-rummager.com

2 SYSTEM-ORIENTED PERFORMANCE MEASURES

The overall retrieval effectiveness can be gauged only if the actual relevancies are known [1]. Let the database $\{x_1, x_2, \cdots, x_i, \cdots, x_N\}$ be a set of N images represented by low or high level features. To retrieve images similar to a query q, each database image x_i is compared with the query image using an appropriate distance function $d(q, x_i)$. The database images are then sorted in a ranked list RL_q according to their distance to the query image such that $d(q, x_i) \leq d(q, x_{i+1})$ holds for each image pair [15].

An important attribute that contributes to evaluating the retrieval system is the Rank(k) index. This index describes the retrieval rank of the k^{th} ground truth image. Consider a query q and assume that the k^{th} ground truth image is found to be the R^{th} result of the retrieval. Then Rank$(k) = R$. Let us note $NG(q)$ the total number of relevant images for the query q.

In [7] some of the most important image retrieval performance measures for a single query image are described. The most commonly used indices which contribute to the formation of performance measures for information retrieval systems are the following[1][7]:

Detections - True Positives: $A_k = \sum\limits_{n=1}^{k} V_n$, where $V_n \in \{0, 1\}$ describes the relevance of the image that appears at position n. If the image belongs to the ground truth of the query then $V_n = 1$, otherwise $V_n = 0$.

False Alarms - False Positives: $B_k = \sum\limits_{n=1}^{k} (1 - V_n) = k - A_k$. This performance measure essentially counts the incorrect results (false positives) that appear in the first k retrieved images.

Misses - False Negative: $C_k = \sum\limits_{n=1}^{N} V_n - A_k = NG(q) - A_k$, where N is the total number of images in the database.

Correct Dismissals - True Negative: $D_k = \sum\limits_{n=1}^{N} (1 - V_n) - B_k$.

By using these indices the following standard information retrieval measures are implemented.

Recall: $R_k = \frac{A_k}{A_k + C_k} = \frac{A_k}{NG(q)} = \dfrac{|\text{retrieved} \cap \text{relevant}|}{|\text{relevant}|}$. Recall essentially describes the ratio of the number of the relevant images within the first k results, to the number of the total relevant images.

Precision: $P_k = \frac{A_k}{A_k + B_k} = \frac{A_k}{k} = \dfrac{|\text{retrieved} \cap \text{relevant}|}{|\text{retrieved}|}$. Precision essentially describes the ratio of the number of the relevant images within the first k results, to the number of the retrieved images.

Recall and precision have often different objectives. If someone wants to see more relevant items (i.e., to increase recall level), usually more nonrelevant ones are also retrieved (i.e., precision decreases) [20]. Each of these two performance measures can be optimized if considered in without the other [21]. For example, we can always achieve a recall value equal to 1, simply by retrieving all the items (the entire database). The precision value in this case decreases dramatically. Thus, precision and recall values have to be used in combination.

Precision absolute value at a given threshold (cut-off) may be precise in many cases, especially during the evaluation of web-based retrieval system. Precision value at a given threshold, e.g. 10 or 20 items, denotes the fraction of relevant items retrieved in these positions. Similarly, recall value at a given threshold determines the ratio between the relevant items retrieved and the number of the relevant items in the database. Recall at small thresholds is not particularly meaningful for queries with many relevant items. Likewise, recall measured at high thresholds seems only of academic importance and is not interesting for users [22].

Generality: $g_k = \frac{A_k}{N}$. It is also known as *Relevant Fraction* and is the fraction of relevant items in a database. Though generality is a major parameter for performance characterization, it is often neglected or ignored [23].

Using these general, standard information retrieval measures as building blocks, one can form the following performance measures [1]:

- Retrieval effectiveness: P_k vs R_k.
- Receiver operating characteristic: A_k vs V_k.
- Relative operating characteristic: A_k vs F_k.
- R-value: P_k at cut-off.
- 3-point average: average P_k at $R_k = 0.2, 0.5, 0.8$.

A commonly used performance measure that combines Precision and Recall is the F-measure, also known as the balanced F-score:

$$F(q) = 2 \times \frac{P_k \times R_k}{P_k + R_k} \tag{1}$$

This formula is also known as the F_1 measure, because recall and precision are evenly weighted. In its more general form, F_β, the F-measure is defined as:

$$F(q) = (1 + \beta)^2 \times \frac{P_k \times R_k}{\beta^2 \times P_k + R_k} \tag{2}$$

Two commonly used F measures are the F_2 ($\beta = 2$) measure, which weights recall higher than precision, and the $F_{0.5}$ ($\beta = 0.5$) measure, which emphasizes precision rather than recall.

Precision and Recall are set-based measures. Therefore, they are considered appropriate for evaluating classification systems but not systems which return ranked lists. In pure classification problems, Precision and Recall, together with the F measure suffice for a complete evaluation of the system.

In the aforementioned problems, ROC graphs [24] are often used for visualizing, organizing and measuring classifiers based on their performance. ROC graphs depict relative trade-offs between benefits and costs (i.e. true positives and false positives). As with any evaluation metric ROC has its limitation, however, placing a classifier in the ROC space gives the observer a fast outlook on its performance with a simplified rule being that a classifier is better than another if it is to the north-west of the first.

Image retrieval systems return rankings and not sets, so we need measures over rankings. In the ROC space, in order to trace an evaluation curve of a ranking classifier, threshold values are used to produce different points in the two-dimensional graph. These thresholds values (strict probabilities or uncalibrated scores) are in fact numeric values that represent the degree of participation of an instance to a class.

In most of the cases, in order evaluate ranked lists, precision-recall curves P_k vs R_k, $(R, P(R))$ are commonly used. Each precision-recall point is computed by calculating the precision at a specified recall cut-off value. For the rest of the recall values, the precision is interpolated. When using the precision-recall curve, one assumes that users choose a rank threshold and only view things above that rank. A very important issue is the definition of this cut-off value. It is common to measure precision at 3 or 11 standard recall levels. Similar to an ROC curve, we can draw thresholds at all ranks and construct precision-recall curves. Then the $(R, P(R))$ curve, together with the total number of images in the database, fully characterize a system which returns a ranking. An obvious drawback of this method is that, two systems may behave differently; one may achieve high precision but low recall, while the other, low precision and high recall. In this case, in the precision-recall space, their curves would intersect and we can't really define which system behaves better. Hence, systems must be evaluated based on the retrieval task. For example, for web-based retrieval systems, where the user is concerned with the relevance of the first results (precision-oriented tasks), without requiring the system to retrieve the entire set of relevant images, the system which achieves high precision is preferable. There are, however, other tasks in which the retrieval of the entire set of relevant items is required. These tasks are known as recall-oriented. Consider, for example, an image retrieval system which retrieves images from patents. The authority which is responsible for the originality of a patent under review is obliged to check all similar patents, and not just the first results. In such tasks, the system which achieves high recall is preferable.

In many cases, in order to compare the performance of different systems, it is desirable to use a single number, which captures the performance of each system instead of a graph. Besides the fact that using a single value is particularly convenient, evidence has shown that it also provides information that in many cases, is not easy to detect in graphs. For example, according to [25], during the first year of ImageCLEF [26,27], a $(R, P(R))$ curve was used to compare different retrieval systems. However, a typical $(R, P(R))$ graph showed similar characteristics of all plotted systems. Thus, in subsequent years, several single value performance measures were employed in evaluating the systems. Image-

CLEF is an initiative to evaluate cross-language image retrieval systems which have been running as part of the Cross Language Evaluation Forum (CLEF). Another advantage of single value performance measures is their intuitive nature. In contrast, an $(R, P(R))$ curve consist of a pair of numbers and, thus, ordinary users cannot quickly interpret what the measure conveys [28].

Single value performance measures are used in order to compare different retrieval systems where most of the retrieval parameters, such as the database, ground truths, and scope are kept constant. As a global estimate of performance using a single value, it is standard to use the average precision (AP).

The average precision for a single query q is the mean over the precision scores at each relevant item:

$$AP(q) = \frac{1}{NG(q)} \sum_{k=1}^{NG(q)} P_q(R_k) \tag{3}$$

where R_k is the recall after the k^{th} relevant image was retrieved. Consequently, the mean average precision (MAP) is the mean of the average precision scores over all queries:

$$MAP = \frac{1}{Q} \sum_{q \in Q} AP(q) \tag{4}$$

where Q is the set of queries q. In the perfect retrieval case $MAP = 1$ and as the number of the nonrelevant images ranked above a retrieved relevant image increases, the MAP approaches the value 0, $MAP \in [0, 1]$. An advantage of the mean average precision is that it contains both precision and recall oriented aspects and is sensitive to the entire ranking.

MAP has been the dominant system-oriented performance measure in information retrieval systems for a number of reasons [29]:

- It has a nice probabilistic interpretation [30].
- It has an underlying theoretical basis as it corresponds to the area under the precision recall curve.
- It can be justified in terms of a simple but moderately plausible user model [31].
- It appears to be highly informative; it predicts other metrics well [32].
- It results in good performance ranking functions when used as objective in learning-to-rank (LTR)[33].

MAP constitutes one of the basic evaluation criteria for the retrieval results in the Text REtrieval Conference (TREC) [34,35], the TrecVid [36] and the ImageCLEF. uses the geometric mean of AP scores.

MPEG-7 [13] [14] proposed a new performance measure, the Averaged Normalized Modified Retrieval Rank (ANMRR) [12]. ANMRR is always in the range of 0 to 1, and the smaller the value of this measure the better the matching quality of the query is. ANMRR is the evaluation criterion used in all of the MPEG-7 color core experiments. Evidence has shown that the ANMRR measure coincides approximately linearly with the results of the subjective evaluation of the retrieval accuracy of search engines [37][12][38]. ANMRR is built using the following indices.

The average rank $AVR(q)$ for a given query q is:

$$AVR(q) = \sum_{k=1}^{NG(q)} \frac{\text{Rank}(k)}{NG(q)} \tag{5}$$

where $NG(q)$ is the number of ground truth images for the query q. If this image is in the first K retrievals then $\text{Rank}(k) = R$ else $\text{Rank}(k) = 1.25 \times K$. K is the top-ranked examined retrievals, where:

$$K = min\left(X \times NG\left(q\right), 2 \times GMT\right) \tag{6}$$

- If $NG(q) > 50$ then $X = 2$ else $X = 4$. Parameter X, as defined by MPEG-7, aims to enable the retrieval systems to have a small number of images in the ground truth.
- $GMT = max\{NG(q)\}$ for all q's of a data set.

The modified retrieval rank is:

$$MRR(q) = AVR(q) - 0.5 \times [1 + NG(q)] \tag{7}$$

The normalized modified retrieval rank is computed as follows:

$$NMRR(q) = \frac{MRR(q)}{1.25 \times K - 0.5 \times [1 + NG(q)]} \tag{8}$$

Finally, the average NMRR over all queries is defined as:

$$ANMRR = \frac{1}{Q} \sum_{q=1}^{Q} NMRR(q) \tag{9}$$

One of the most significant advantages of ANMRR is that, similar to MAP, it combines both precision and recall oriented aspects. ANMRR has already been used by several image retrieval systems [39][40].

The authors in [41] demonstrate how the evaluation results depend on the particular content of the database. For the same ground truth, the performances of the systems degrade for larger image databases. All the above retrieval performance measures do not take into account the size of the image database. Thus, the performance of a scaled-up version of an image retrieval system cannot be predicted.

Huijsmans and Sebe [42] [23] highlighted this limitations on the typical precision-recall curves and proposed additional performance measures to overcome these limitations. They proposed the use of generality along with precision and recall parameters. The result is a three-dimensional representation, which can be reduced to a two-dimensional graph by keeping constant one of the parameters. Therefore, the graph plots precision vs recall on the y-axis against generality on the x-axis.

A measure that takes into consideration the database size is the Normalized Averaged Rank (NAR) proposed in [7]. Using the definition from [43], NAR is defined as:

$$\mathrm{NAR} = \frac{1}{N \times NG(q)} \left[\sum_{i=1}^{NG(q)} \mathrm{Rank}(i) - \sum_{i=1}^{NG(q)} (i) \right] \tag{10}$$

This measure is 0 for perfect retrieval, and approaches 1 as performance worsens. NAR is basically a complement of the normalized recall proposed in [44]. The average NAR over all queries is defined as:

$$\mathrm{ANAR} = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{NAR} \tag{11}$$

All the aforementioned evaluation measures consider the retrieved data as either relevant or non-relevant to the query. Even though the matter is not investigated in the current work, it is important to mention that the concept of non-binary relevance is much employed in recent evaluation approaches. Assume for example the case in which the ranking list of a system is: $RL_1 = X_1, X_2, X_3, X_4, X_5$. At the same time, a second system produces the following ranking list: $RL_2 = X_2, X_3, X_1, X_4, X_5$. We also assume that X_1, X_2, X_3 are relevant with the query image. In both cases, e.g., AP=1 and NMRR= 0. If the images had a different level of relevance, the ranking order would be a much more important factor. Highly relevant documents are more useful when appearing earlier in a search engine result list and highly relevant documents are more useful than marginally relevant documents, which are in turn more useful than irrelevant documents.

3 PERFORMANCE STUDY OF MAP AND ANMRR

As mentioned in Section 2, the most widespread image retrieval performance measures with the ability to evaluate the systems using only one number are AP (Average Precision) and NMRR (Normalized Modified Retrieval Rank). At [45] NMRR is used to measure the performance of a set of descriptors for natural images while at [15], AP is used for the same databases. At [15] and [46] AP is used to measure the performance of descriptors for medical images. It can be observed, however, that there are deviations between the results of these two techniques. In order to make it easier to compare the results, $1 - \mathrm{AP}$ shall be used so that in both performance measures, perfect retrieval will produce a 0, while as more non-relevant images retrieved appear in the results, both performance measures approach a value of 1. Indicatively, we can mention the results of the Color and Edge Directivity Descriptor (CEDD) [47] in the Wang [8] database, where at the performed experiment, the queries and their ground truth given at [40] were used. In this case ANMRR is equal to 0.2528 while $1-\mathrm{MAP}$ is equal to 0.4109. It is apparent that these values differ significantly, giving quite different evaluation score to a retrieval method.

In order to have a better look in the way these performance measures operate and address the issue of their significant deviation, we utilized an oversimplified Know-Item example. We employed an artificially generated database with 20 images ($N = 20$). The experiments that follow serve purely as an illustrative tool in order to examine the behavior of MAP and NMRR, since the artificially generated database of 20 images can by no means be a credible set for retrieval purposes. An image from the database was selected to be the query image and its ground truth was taken to be the image itself ($NG(q) = 1$). Following this, the effectiveness of both $1 - \mathrm{AP}$ and NMRR was estimated, both for those scenarios in which the query image is retrieved consecutively from position 1 to 20. Figure 1 presents the results when $\mathrm{Rank}(q)$ take values in the range 1 to 20. The horizontal axis depicts each position where the image was retrieved, while the vertical axis corresponds to the values for $1-\mathrm{AP}$ and the NMRR.

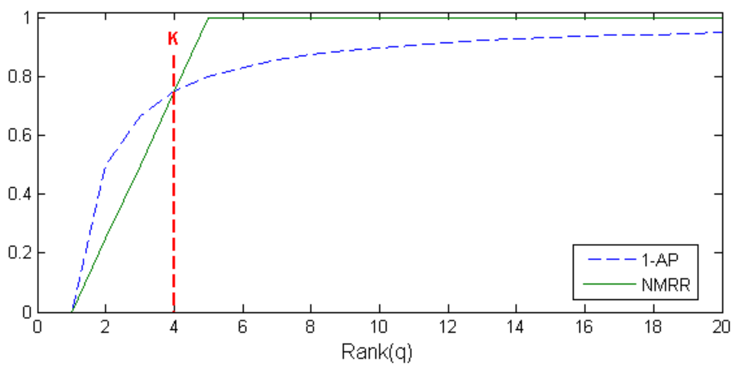

Fig. 1 Results of $1 - \mathrm{AP}$ and NMRR for $NG(q) = 1$, $N = 20$

Observing the results of Figure 1, the following conclusions are drawn. The graphical representation of $1-\mathrm{AP}$ appears to be non-linear where its gradient is larger in the first $\mathrm{Rank}(q)$ values and then becomes gradually smaller. In the first K (see Figure 1) $\mathrm{Rank}(q)$ positions, $1 - \mathrm{AP}$ appears stricter than NMRR because it takes larger values and therefore characterizes the retrieved results as less relevant. This result is to be expected, given that AP, and by extension $1 - \mathrm{AP}$ has a natural top-heavy bias. On the other hand, NMRR appears to be stricter than $1 - \mathrm{AP}$ and seems to "punish" the system when $\mathrm{Rank}(q) > K$. This behavior can be easily explained if one takes into account the assumption made during NMRR formation. According to this assumption, if an image appears after the position $K = \min\,(X \times NG(q), 2 \times GMT)$ then this image is considered as not retrieved. That's why NMRR is equal to 1 for all the Rank(q) $> (K + 1)$.

$$NMRR(q) = 1, \forall \text{Rank}(q) > (K + 1) \tag{12}$$

In contrast, $1 - \text{AP}$ considers that each image contributes to the retrieval evaluation process for each Rank(q).

Moreover, it can be observed that NMRR is composed of three consecutive linear functions. It increases linearly from position 0 to K with a gradient of α, it increases from point K to $K + 1$ with a gradient of β (when $NG(q) = 1$ the two gradients are equal) and from position $K + 1$ it becomes a straight horizontal line with NMRR being always equal to 1.

In order to see how these 2 retrieval evaluation behave in more complex scenarios, we utilize a second example, in which we take each query image q to include 2 images in its ground truth ($NG(q) = 2$). These images are defined as j and i. Similar to the first example, the testing database contains 20 images.

We study the effectiveness of the retrieval system when image i was retrieved in position Rank(i), while image j was retrieved in position Rank(j), where Rank(j) $\in [1, \text{Rank}(i) - 1]$. In order to test all possible combinations of Rank(i) and Rank(j) we employed the following pseudo code:

```
Combined_Rank=0;

For (int i=2; i=20; i++)
{
 For (int j=1; j=i-1; j++)
 {
  Rank(i)=i;
  Rank(j)=j;
  Combined_Rank++;
 }
}
```

This pseudo code, for each combination of Rank(i) and Rank(j), generates a unique identification, the *Combined*_Rank, which includes information on both the position of image i, as well as the position of image j. In total, 190 ordering combinations are tested.

For each combination, the $1 - \text{AP}$ and NMRR are calculated, resulting in the performance shown in Figure 2. The horizontal axis describes each *Combined*_Rank while the vertical axis displays the values for $1 - \text{AP}$ and NMRR.

In order to reach more solid conclusions, we depicted in Figure 3 the three-dimensional representations of the results for $1 - \text{AP}$ and NMRR for every ordering combination. The 2 axis which shape the plane describe Rank(i) and Rank(j) while the vertical axis displays the values of $1 - \text{AP}$ and NMRR.

The projection of the 3-D graphs on 2-D graphs (see Figure 4) where the horizontal axis is Rank(i) and the vertical axis corresponds to $1 - \text{AP}$ and

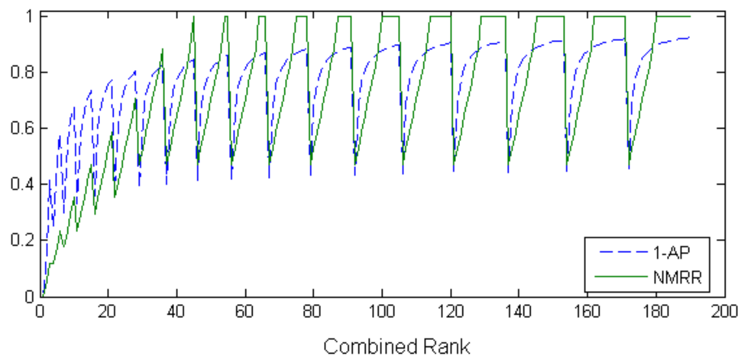

Fig. 2 Results of $1 - \text{AP}$ and NMRR for $NG(q) = 2$, $N = 20$

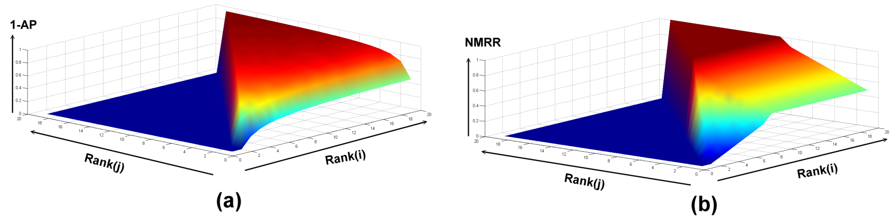

Fig. 3 3D Representation of the results of (a) 1 - AP (b) NMRR for $NG(q) = 2$, $N = 20$

NMRR respectively, depicts two curves each one representing the best and worst (j, i) combination order. Figure 4(a) shows the curves for $1 - \text{AP}$ while Figure 4(b) shows the two curves for NMRR.

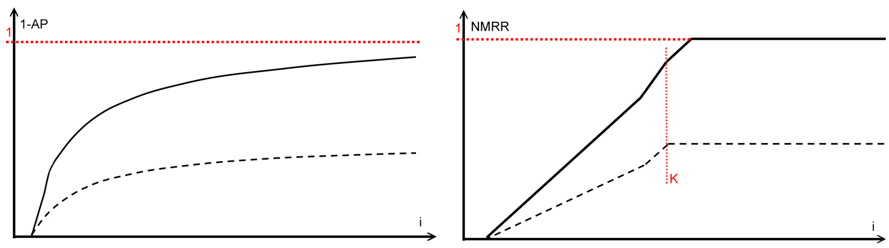

Fig. 4 Curves forming the (a) $1 - \text{AP}$ and (b) NMRR values for $NG(q) = 2$, $N = 20$

The horizontal axis of the two curves describes the position in which image i appears while the vertical axis describes the retrieval performance. In both Figures 4(a) and (b), the lower curve describes the retrieval success rate if image i was retrieved in the position Rank(i) while image j was retrieved in the position Rank$(j) = 1$. Thus, it describes system effectiveness, if the one image can be retrieved first in the ranked list while the second in position i. As Rank(j) increases, while i remains constant, the value of both $1 - \text{AP}$ and NMRR approaches the lower curve. In the worst case, where image i has retrieved in the position Rank(i) and image j has retrieved in the position Rank(j) = Rank(i) − 1, the performance of the systems is described by the upper curves.

Essentially, the upper curve displays how much the precision affects each method, while the lower curve shows the contribution of recall. Looking at the $1 - \text{AP}$ curves, we can observe that, if all the results of ground truth are retrieved in early positions, that is, with a small Rank(i), the value of $1-\text{AP}$ is much higher than the equivalent value of NMRR, lending credence to the conclusion that $1-\text{AP}$ is much more oriented towards early precision results than ANMRR. However, as the value of Rank(j) increases, and therefore the value of early precision decreases, the value of $1 - \text{AP}$ show a significant increase.

The manner in which recall and precision information are connected to the NMRR is similar to that in $1 - \text{AP}$. In the first steps, i.e. for small Rank(i), the value of NMRR is smaller than the corresponding $1-\text{AP}$ value. The main difference, however, appears after position K, where it is obvious that the lower curve, yields greater values than those for $1 - \text{AP}$. A similar behavior is shown in the upper curve, with the precision parameter playing a basic role so that the system is not graded with the worst possible score. By observing the graph we see that for $min(\text{Rank}(i), \text{Rank}(j)) > K$ we have NMRR=1. For the same Rank(i) and Rank(j) positions, $1 - \text{AP}$ grades the system with a much smaller value. In the case where $NG(q)$ is greater than 2, the operating principle of both $1 - \text{AP}$ and NMRR remains the same.

Having studied the behavior of these two performance measures, we can draw the following conclusions. The biggest distinction between these two measures is related to how they treat early positions (low-ranking results). AP is consistently correlated with the first 10 to 20 positions, while NMRR increases linearly from the first to the Kth position. The K position is dynamically calculated for each query and is related to the number of the relevant items. As mentioned in the Introduction, we argue that, the evaluation of content-based image retrieval systems, must take into account the specificities of the results. Due to the nature of the low-level features that CBIR systems use, images that are visually similar but semantically distinct from the query often appear among the early retrieval positions. Additionally, the fact that the results of an image retrieval system are often viewed in table of images on the screen and not in a list as text results are, enhance the observation that the performance measures, which evaluate CBIR systems, should not be influenced only by the results in the first N positions. A more preferable approach is to use a threshold which will be directly connected, either with the generality of the query, or with the number of relevant to the query items.

NMRR, which was proposed for use predominantly in image retrieval systems, corresponds to the goals of the CBIR system's results and gives a bias to the recall at K. In other words, NMRR is evaluating the capability of the system to retrieve, in the first K positions, as many results as possible from the ground truth. Systems which retrieve results after these first K positions, are ranked with very high values. On the other hand, AP gives weight to early precision during results evaluation, which in effect highlights the capability of the system to retrieve as many results as possible in the early positions. This implies that, especially for queries with a small ground truth, AP 'punishes' the retrieval system even if the images appear in a relatively small Rank(k).

Additionally, even though NMRR was designed to evaluate image retrieval systems, the adopted assumption, that if the image appears after the K^{th} position it is considered as not having been retrieved, seems to be problematic. The principle of operation of NMRR does not allow a comprehensive evaluation of recall-oriented tasks since it completely ignores the position in which each image eventually appears. As shown in Figure (4)(b), from position $K+1$ there is no information about the ranks at which relevant items are retrieved. Assume for example two image retrieval systems T_1 and T_2, a query Q, $NG(Q) = 2$ and a database size equal to N. Both systems are retrieving the first relevant image in the first position. T_1 retrieves the second relevant image in position 100, while T_2 retrieves the second relevant image in position 1000. In a recall-oriented task, system T_1 has a clear advantage over the system T_2. Under ANMRR, however, the systems perform equivalently.

In comparison, even though MAP is not the most appropriate method for recall-oriented tasks [48], it still carries information about the rank of all the relevant items. One, however, should keep in mind that during the evaluation of a recall-oriented system, it is important for a performance measure to take into account not only the recall value, but also the ranks at which the relevant items are retrieved [48].

A common disadvantage of both methods is that they do not take into account the generality of the queries and thus they can not predict the behavior of a scaled-up version of the system. Experimental results in Section 5.2 demonstrate the effects of this drawback.

4 MEAN NORMALIZED RETRIEVAL ORDER

The conclusions drawn in the previews sections concerning NMRR and $1-\mathrm{AP}$ lead us in defining a set of properties of a new performance measure. Such a measure should evaluate the retrieval systems by taking into account the position where each image appears, even if it is retrieved in positions which the web-based/precision oriented systems would have rejected. Thus, the new performance measure must be differentiated from NMRR with respect to the parameter which determines that if an image is retrieved after position K, it is considered as non-retrieved. In the proposed performance measure an upper limit will also be defined. However, this upper limit is now dynamically designated for each query by taking into account the generality of the query. Furthermore, the images retrieved after this limit will still contribute to the performance measure but at a lower degree. Using this approach, the new performance measure can predict the behavior of a scaled-up version of the system. Moreover, this new performance measure, unlike AP, must not be biased on the top-10 or top-20 results. Rather, it should take into account the specificities of the results of a CBIR system, as well as the fact that the results of an image retrieval system are often viewed in a table of images on the screen and not in a list as text results are.

The **Gompertz Sigmoid Function(GSF)** [49] does satisfy these conditions. GSF is a mathematical model for a time series, where growth is slowest at the start and end of a time period. Originally formulated in 1825 to model the mortality rate of a population, it later became one of the most frequently used laws to describe tumour growth (it is currently applied in other contexts, both in biology and in economics)[50]. The general form of this function is:

$$f(t) = ae^{be^{ct}} \tag{13}$$

parameter a controls the amplitude of the function and parameters b and c are always negative real numbers. Given that we want the function to take values in the range of $[0, 1]$, we set $a = 1$.

The combination of parameters b and c determines the point at which the function approaches the value 1 as well as its gradient. In order to calculate parameters b and c we make the following assumptions:

1. If an image is retrieved at position K, where K is dynamically calculated for each query and depends upon the size of its ground truth then the Normalized Retrieval Order (NRO) is equal to 0.95.
2. If an image is retrieved at position $\frac{K}{2}$ then the Normalized Retrieval Order (NRO) is equal to 0.50.

According to ANMRR, K is defined as: $K = \min$ ($X \times NG(q)$, $2 \times GMT$), $X = 2$ when $NG(q) > 50$ else $X = 4$. The proposed method method uses the query generality $g(q)$ to define the K position as:

$$K = \begin{cases} 4 \times NG(q) & g(q) \geq 0.01 \\ F[g(q)] \times NG(q) & g(q) < 0.01 \end{cases} \tag{14}$$

where

$$F[g(q)] = \frac{0.04}{g(q)} \times NG(q) \tag{15}$$

In other words, if the query generality is higher than a given value, then we adopt the NMRR assumption, $(K = K)$. But when the generality is smaller, the position K increases linearly.

Under these assumptions, solving Eq. 13 leads to $b = -9.3668$ and $c = -5.2074/K$. Therefore, the Normalized Retrieval Order for each image retrieved at position Rank(k) is equal to:

$$NRO(q) = \begin{cases} 0 & , \frac{k}{\text{Rank}(k)} = 1 \\ e^{-9.3668 \times e^{-5.2074 \times ARANK(k)}} & \frac{k}{\text{Rank}(k)} < 1 \end{cases} \tag{16}$$

where

$$\text{ARANK(k)} = \frac{\text{Rank}(k) - 1}{K - 1} \tag{17}$$

We repeated the Known-Item example of Section 3, and used an artificially generated database with 20 images ($N = 20$). As query image, an image was selected from the database. The corresponding ground truth was the image itself ($NG(q) = 1$). We then calculated the effectiveness of the proposed performance measure, for those scenarios in which the query image is retrieved consecutively from position 1 to 20. Figure 5 presents the results when Rank(q) takes values in the range 1 to 20. The horizontal axis shows the specific location in which the image was retrieved, while the vertical axis shows the values for NRO. In the same graph the corresponding NMRR and $1 - \text{AP}$ values are also depicted.

As Figure 5 shows, in the first results the gradient of the NRO is smaller than the gradients of $1 - \text{AP}$ and NMRR. This indicates that the proposed performance measure is less biased towards early precision than the other 2 measures. From position K onwards, beginning with value 0.95, the NRO increases with a very small gradient, approaching the value 1. We can therefore conclude that NRO is more advantageous than NMRR since it is in a position to accurately evaluate each specific retrieval location, even after the first K positions.

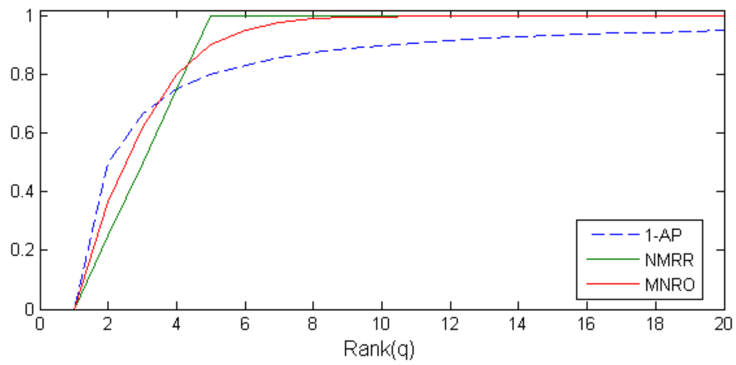

Fig. 5 Results of $1 - \text{AP}$, NMRR and MNRO for $NG(q) = 1$, $N = 20$

If the ground truth of the query q contains more than one image then the Mean NRO(q) is calculated as:

$$MNRO(q) = \frac{1}{NG(q)} \sum_{k=1}^{NG(q)} NRO(k) \tag{18}$$

Next, we repeated the experiment, increasing the size of the database. Figure 6 illustrates the behavior of the MNRO for a query with a single relevant image over four different databases. The first database consist of 100 images, the second one contains 1000 images, the third one 10000 images and finally the fourth one includes one million images. Please note that we assume that all the images in the databases are embedding images[23] (irrelevant to the query images) and in each database, only one is considered as relevant to the query.

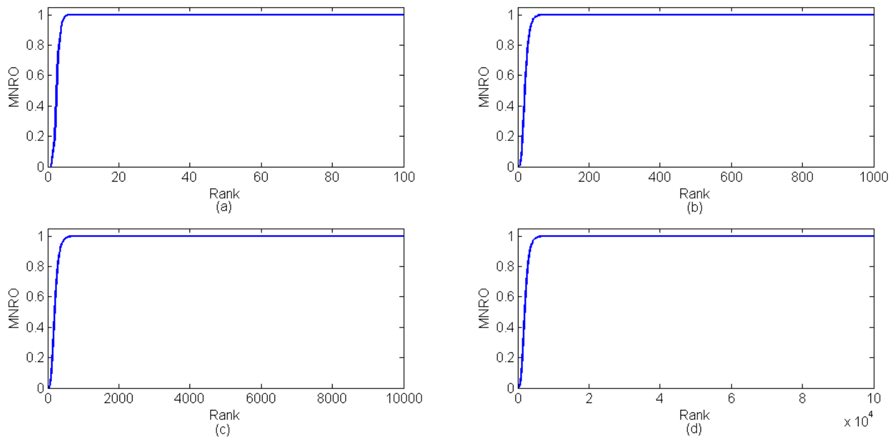

Fig. 6 Results of MNRO for $NG(q) = 1$, (a) $N = 1000$, (b) $N = 10000$, (c) $N = 100000$ and (d) $N = 1000000$

As one can see, the $F[g(q)]$ factor gives the capability to MNRO to adjust itself in order to keep the same behavior over different database sizes. This property gives the ability to the proposed performance measure to adjust according to the generality of the query. The assumption behind the $F[g(q)]$ is based on [23] and [7], which argues that the number of non relevant items retrieved is linearly correlated with the size of the database. The experimental results presented in Section 5.2 confirm this argument.

In our next evaluation, we repeated the example of Section 3 in which the ground truth of a query image consist of two images, j and i. All the possible order combinations of the images are tested according to the pseudocode of Section 3. The results are shown in Figure 7. In the same graph we depict the relevant values from NMRR and $1-\text{AP}$. Even in this case one can observe that the MNRO satisfies its design requirements. Its gradient in the first results is smaller than the gradient of $1-\text{AP}$ and it is capable of evaluating each retrieved image, without disregarding any images.

Similarly to Section 3, Figure 8 provides the 3-dimensional representation of the results for MNRO for every ordering combination. The 2 axes which form the horizontal plane correspond to Rank(i) and Rank(j), while the vertical axis depicts the MNRO values.

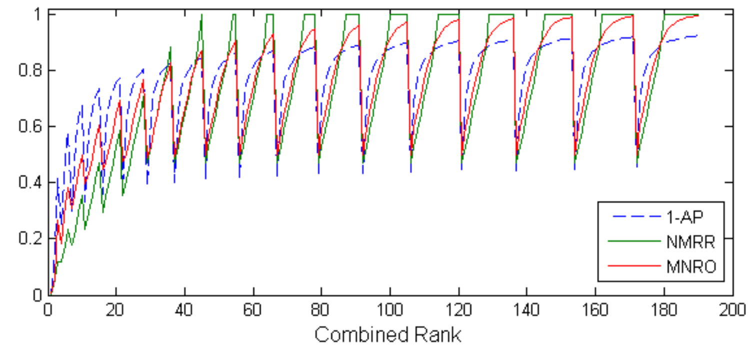

Fig. 7 Results of $1 - \mathrm{AP}$, NMRR and MNRO for $NG(q) = 2$, $N = 20$

By observing this graph it is easy to distinguish the 2 curves which shape the influence curve for precision and the contribution curve for recall, exactly as in the case for NMRR and $1 - \mathrm{AP}$ illustrated in Figure 4. It can be seen that the performance measure is oriented towards the first K results. Systems which present their results in positions after position K, are evaluated with very high values. The larger the number of results which appear after this position, the higher the value returned by the system.

In the early results, the value of MNRO is definitely smaller than the equivalent values of $1 - \mathrm{AP}$, and approximately at the levels of the values for NMRR. After position K the lower curve yields larger values than the corresponding ones for $1 - \mathrm{AP}$, and even in this case, the values are at similar levels to the corresponding ones for NMRR. However, in the event that $min(\mathrm{Rank}(i), \mathrm{Rank}(j)) > K$, where NMRR=1, the values for MNRO increase linearly with a very small gradient, approaching a value of 1, without however ever becoming equal to a value of 1. In the corresponding positions, the value of $1 - \mathrm{AP}$ is definitely smaller.

To improve the readability of Figure 8, we marked the enveloping curves as A and B. Curve A describes the MNRO value for the best case scenario, in which the first relevant image is retrieved in position Rank(j) while the second relevant image is retrieved in position Rank(i)= Rank(j)+1. Curve B, on the other hand, describes the worst case scenario, in which, the first relevant image is retrieved in position Rank(j), while the second relevant retrieved in position Rank(i)= N.

In the case of perfect retrieval $MNRO(q) = 0$, while as the rank errors increase, the MNRO approaches the value 1, $MNRO(q) \in [0, 1]$. Finally, the average retrieval rank over all queries is defined as:

$$AMNRO = \frac{1}{Q} \sum_{q=1}^{Q} MNRO(q) \tag{19}$$

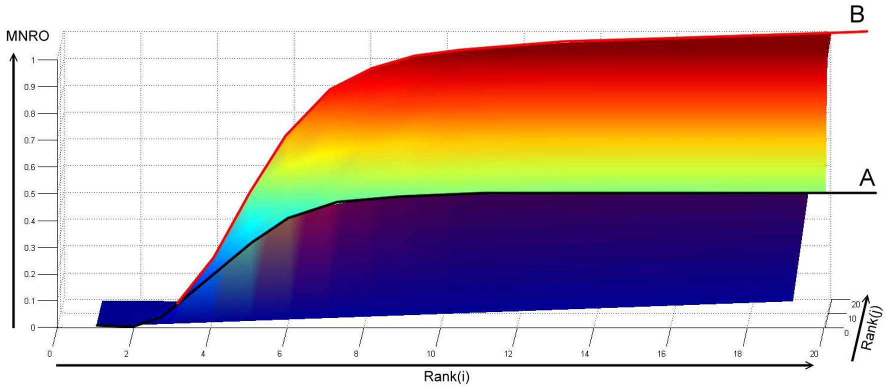

Fig. 8 3D representation of the MNRO results for $NG(q) = 2$, $N = 20$

The proposed retrieval rank performance measure, like ANMRR and MAP, offers the capability to evaluate a system on the basis of only a single value, which includes information about both precision and recall.

5 EXPERIMENTAL RESULTS

Before presenting the experimental results, it is very important to review the attributes of a good performance measure. First and foremost, we believe that a performance measure should be easy to interpret. Using the curves introduced in Section 4, one can easily analyze the behavior of the proposed performance measure.

A performance measure should also separate well good from poor techniques. In Section 5.1 we evaluate the MNRO and highlight its advantages over NMRR and AP on a small database by presenting the evaluation results of the three performance measures on different ranked lists.

Moreover, we consider that a good performance measure should provide consistent results, especially over systems with different generality. In Section 5.2, a second experimental setup evaluates the ability of the proposed performance to take into account the generality of the queries during the retrieval procedure. The experiments demonstrate the consistency of the results we obtained when using the proposed performance measure.

Finally, we believe that it is very important for a performance measure to correspond to human perception. Thus, in the third experimental setup, described in Section 5.3, subjective evaluation by human users is taken into account. For the same database size and the same ground truth size for query q, we randomly create 50 different ranked lists and do a case study employing 30 users. Experimental results demonstrate that the proposed performance measure is closer to the user preferences than other performance measures.

In order to further encourage researchers and practicioners to use the proposed performance measure we show, in Section 5.4, the performance of the proposed performance measure in actual retrieval scenarios on three well-known benchmarking databases. The retrieval is performed using several low level features from the literature. We evaluate the results using AMNRO, ANMRR, AP, P(10) and P(20), where P(10) and P(20) denote the precision at the first 10 and 20 results respectively.

5.1 Evaluating the ranked lists

Figure 9 illustrates the hypothetical results produced by the retrieval of a query q with $NG(q) = 5$. Each retrieval result is associated with a hypothetical ranked list. For example in the ranked list 'A' the '+' symbols describe that the 5 ground truth images were the first 5 retrieved images. On the other hand, the ranked list 'E', with its corresponding '+' symbols indicates that the five ground truth images were retrieved as the 1st, 2nd, 3rd, 40th and 41st image respectively. In all the cases, $N = 100$. Table 1 presents the values of the NMRR, AP and, MNOR. In the same table the ranked lists are presented, as it was formed according to the values of each performance measure.

Note once more that, the value of the NMRR(q) and the MNRO(q) is 0 with perfect retrieval while for the AP(q) it is 1.

A	1	2	3	4	5	6
	+	+	+	+	+	

B	1	2	3	4	5	6
	+		+	+	+	+

C	1	2	3	4	...	100
	+	+	+	+		+

D	1	2	3	...	30	31
	+	+	+		+	+

E	1	2	3	...	40	41
	+	+	+		+	+

Fig. 9 Hypothetical Retrieval Results

The following conclusions can be drawn from the results. In example A, where all the images were correctly retrieved, ANMRR=ANMRO=0 and AP=1. The advantages of the proposed performance measure over AP can be more clearly seen in examples B and C. In example B, we observe that a single false alarm was detected in position 2. At the same time, in example C, in order to retrieve all images from the ground truth, it was required to retrieved a total of 100 images. This means, that the last relevant image was retrieved last from the data. In both these cases, AP evaluates the system with exactly the same value $\mathrm{AP}(q_B) = \mathrm{AP}(q_C) = 0.8100$.

These results confirm the fact that AP is oriented towards favouring early results. Moreover, the single false alarm (non relevant retrieved image) in position 2 (example B) gets the same penalty as in example C where the fifth ground truth image is retrieved after the entire database is retrieved. The proposed performance measure evaluates the results in example B with a value at a level fairly close to perfect retrieval score, $MNROR(q_B) = 0.0314$, which is quite close to the corresponding value given by NMRR.

In example C, the proposed performance measure evaluates the system with a value in the same order of magnitude with that given by AP and NMRR, penalizing the retrieval system for its bad performance in the retrieval of the 5th ground truth image.

Experiment	AP(q)	Rank	NMRR(q)	Rank	MNRO(q)	Rank
A	1.0000	1	0.0000	1	0.0000	1
B	0.8100	2	0.0364	2	0.0314	2
C	0.8100	2	0.1818	3	0.2000	3
D	0.6589	4	0.3727	4	0.3988	4
E	0.6444	5	0.3727	4	0.3999	5

Table 1 Experimental Results

Examples D and E show the advantages of the proposed performance measure against NMRR. In example D, 3 relevant results were retrieved at the first 3 positions and were followed by 26 non-relevant items before the appearance of the remaining 2 relevant results in positions 30 and 31. On the other hand, in example E we have the retrieval of the first 3 relevant images in the first positions, we then however require 10 more non-relevant images in order to retrieve the entire relevance set. In both examples, the NMRR value is the same, $NMRR(q_D) = NMRR(q_E) = 0.3727$, because according to NMRR if a retrieved ground truth image appears after the 20th position it is considered as non retrieved. On the other hand, the proposed performance measure if able to merit the differences of the ranked lists, evaluating example D with $MNRO(q_D) = 0.3988$ and example E with $MNRO(q_E) = 0.3999$.

An additional point is that, the scores of the proposed performance measure for examples D and E are greater than the scores for example C. This occurs because the proposed measure penalizes with greater values those systems that retrieve relevant images after the K^{th} position. The more images retrieved after this position, the greater the value of MNRO.

In conclusion, the experimental results indicate that the proposed measure is less oriented towards early results. At the same time, it is capable of continuing the evaluation of the retrieval systems, even if these retrieve results after position K.

5.2 Query Generality

In order to evaluate the ability of the proposed performance measure to take into account the generality of the queries during a retrieval procedure, we employed the ImageCLEF 2010 Wikipedia collection data. This database consist of 237,434 images, associated with noisy and incomplete user-supplied textual annotations and the Wikipedia articles containing the images. There are 70 test topics, each one consisting of a textual and a visual part. The details of the creation of this database, including research objectives, data collection etc., are provided in the overview paper [26].

In our experiment, we created 3 sub-sets of images from the database and we chose 20 queries. The first sub-set consist of 77,300 images. In the second sub-set 77,300 additional images were used, for a total of 154,600 images. The third set contains the entire dataset. It is very important to note that all the relevant to the queries images are included in the first sub-set (and hence the 2nd and 3rd sub-sets as well).

The query images themselves are not part of the database, making the experiment more realistic. In most of academic settings, query images are part of the database. This, however, potentially influences the results since the query image itself is often in the first position, biasing the results, especially in the case where MAP is employed.

Each query consist of a single image. We index the database and the queries with Color and Edge Directivity Descriptor (CEDD)[45]. We evaluate the results using AMNRO, ANMRR, MAP as well as with NAR. The experimental results are presented in Table 2.

Set	MAP	Dev.	ANMRR	Dev.	NAR	Dev.	AMNRO	Dev.
A	0.0375		0.9202		0.2843		0.8356	
B	0.0237	36.8%	0.9457	2.77%	0.2859	0.56%	0.8368	**0.14**%
C	0.0184	50.9%	0.9574	4.04%	0.2873	1.05%	0.8360	**0.05**%

Table 2 Investigating the Generality Independence Ability

We define the value obtained by each performance measure at the sub-set A as baseline. For each sub-set, we calculate the percentage difference of the result from the baseline. As one can see in Table 2, MAP presents the highest percentage deviation among the other performance measures reinforcing the conclusion that it can not adjust to changes in the database size. To investigate the reason of this deviation, we present the P(10) results for the 3 sub-sets: $\mathrm{P}(10)_A = 0.1600$, $\mathrm{P}(10)_B = 0.1300$ and $\mathrm{P}(10)_C = 0.1000$. Translating the numbers, we can observe that in first sub-set, on average, 1.6 out of 10 images on the first positions were relevant. On the other hand, on the third sub-set only 1 out of 10 results were relevant. These results give further credence to the observation that MAP is highly correlated to the early positions. Increasing the number of the non relevant images in the early positions contribute to the decrease of MAP.

The deviation of the ANMRR values is related to the fact that the position K, which determines the bias of the performance measure, considers only the size of the ground truth, without taking into consideration the size of database. Normalized Average Rank (NAR) seems to be more stable than the other two performance measures. NAR assumes that the number of non-relevant items retrieved is linearly correlated with the size of the database. This postulate makes NAR a generality-independent performance measure.

AMNRO, seems to outperform all the other performance measures in terms of the ability to take into account the generality of the queries during the retrieval procedure. The reason relies on the fact that K employ information about the database size as well as about the number of the relevant images. The deviation between the first 2 sub-sets is **0.14**% while the deviation between the first and the third sub-sets is **0.05**%.

5.3 Comparisons to human evaluation

In order to determine which of the 3 retrieval performance measures is closer to human perception, we conducted the following experiment.

Thirty individuals, students of the Electrical and Computer Engineering Department of the Democritus University of Thrace, Greece, most of which were members of the DUTH's Robotic Team[2], participated in an electronic survey. More detailed information on the participants of the survey can be found in Table 3.

To facilitate the electronic survey, a software application was built.Each user, after entering some personal data, is asked to answer 10 questions. To complete the process, each user must answer all the questions. In each question, a set of 5 ranked lists (A, B, C, D, E) appears. Please note that the ranked lists does not contain images, but single numbers. Each number corresponds to the position in which a relevant image retrieved. For example, the ranked list A, consist of the numbers 33, 38, 39, 83 and 97. This mean that the first relevant image retrieved at the position 33, the second relevant image retrieved in position 38 etc.. The ranked lists sets are randomly produced, but once they are produced they remain fixed and are the **same for all users**. Next to each ranked list the values of $1 - AP$, NMRR and MNRO appear, under the labels "Method1", "Method2" and "Method3". The correspondence between the performance measures and the pseudo labels changes randomly for each question. Therefore, the user can not guess the correspondence. In each set, the order of appearance of the values changes randomly. At the same time, the form shows the order in which the ranked lists are ranked with each retrieval performance measure. As in Table 1, the ranking order shows which of all the ranked list of the set exhibits the best behavior.

For each of the sets the user is called to vote (*select*) which of the 3 ranks, as derived from each of the 3 performance measures, more closely matches

[2] http://www.ee.duth.gr/acsl/duthrobotics/index.html

his/her own ranking. Moreover, the user has the option to disagree with all the rankings shown, and to suggest his own ranking using the last column "Custom Ranking" to enter his scores. Additionally, the user is also given the choice to select more than one ranks as most appropriate, in case of ties. The process is repeated for all 10 sets.

People Participating in the Survey	30
Questions Answered By Each User	10
Average Age	22
Standard Deviation to the Age	1.3870
Educational Level	Students
Average Time for Filling in the Questionnaire	18 min.
Standard Deviation to the Time	4.6710

Table 3 Survey ID

In order for the participants to get a feeling of what they are evaluating, the following scenario is told. "There are 5 web-based image retrieval systems. Each system accepts a query (an input facial image) and after searching a database returns facial images. It is assumed that for each query image the database always contains a depository of 5 similar to the query image. The retrieval results of these 5 systems appear to be the respective ranked lists (A, B, C, D, E) appearing in each question". Judging from the position of appearance of the relevant images in each ranked list the users are called to rank each retrieval system (each ranked list) and to determine whether they agree with one of the rankings given by the three pseudo-labeled methods or they prefer to give their own ranking.

Even though the participants are students of the Electrical and Computer Engineering Department, they are not familiar with the image retrieval procedure. We assume that in a more realistic scenario, where images rather than ranking lists were used, the results of the users would be biased by the similarity between the query and a result. For a relevant item retrieved in a specific position, two different users might evaluate the system in different ways. We tried to reduce the subjectivity of the results on how people evaluate ranked lists and not on how they judge how relevant is a result. All three performance measures we employed are using the binary relevance assumption. Additionally, by incorporating facial images, we are trying to achieve a trade-off between precision-oriented and recall-oriented tasks. We assume that, if someone searches for facial images on the web, especially for personal facial images, he/she is concerned with how many images will appear in early positions and with retrieving all available online images.

The answers of the participants for each set of ranked lists are summarized in Table 4, where each number denotes the number of individuals that agree with the ranking of the particular performance measure. Column "OTHER" contains the number of participants who preferred their own ranking. It is apparent that the proposed performance measure was selected by the majority

	AP	NMRR	MNRO	OTHER	Participant's Choice
Set 1	5	9	13	3	**MNRO**
Set 2	8	10	12	0	**MNRO**
Set 3	20	6	20	4	**MNRO**-AP
Set 4	8	20	20	2	**MNRO**-NMRR
Set 5	8	10	10	2	**MNRO**-NMRR
Set 6	10	4	14	2	**MNRO**
Set 7	7	10	11	2	**MNRO**
Set 8	6	9	13	2	**MNRO**
Set 9	14	14	14	2	**MNRO**-AP-NMRR
Set 10	4	17	8	1	NMRR
Total Votes	90	109	**135**	20	

Table 4 Votes Per Set

of users in almost all the sets, collecting in total 135 votes. In some sets, the sum of the votes exceeds 30, which is the total number of participants. The reason for this is, that in some ranked lists, there were ties. In set 3 and set 9, there is a tie between the values of AP and MNRO, while in set 4, there is a tie between NMRR and MNRO.

Percentage-wise, we see that AP was the participant's choice 25.42% of the times, NMRR 30.79% and the MNRO **38.14%**. Moreover, a 5.65% declared that they did not agree with any of the choices.

These results, may confirm the conclusions drawn in [12][38], which state that there is a high correlation between NMRR and the retrieval quality explored in subjective experiments. This correlation is further strengthened in MNRO. NMRR exceeds AP in votes, in 7 of the 10 sets. AP is in first place in only 2 sets, in which however, it is tied with MNRO. The proposed performance measure gains first place in participants selection in 90% of the sets, losing only in set 10 from NMRR.

We assume that the proposed performance measurement was selected by the majority of the participants mainly due to the common way that a human judge and our method deal with non-relevant results in the early positions. The task we chose is purely an image retrieval task. Although we noted that the participants are not familiar with the image retrieval procedure, we can only assume that they have great experience with the way web based image retrieval engines present their results. Thus, we consider that the participants evaluate the results of the survey based on criteria related to this experience. As we stressed earlier, the results of a web based image retrieval engine, are often viewed in table of images on the screen and not in a list as text results are. People that are used to this kind of result depiction tend to evaluate the results less strict based on the absolute rank position.

The aforementioned assessment also justifies the fact that the NMRR is the second choice of participants while the early-precision-oriented MAP method comes last in the people's choice. The criterion that mainly contributed in the precedence of the MNRO over the NMRR is related to the way the retrieval results are evaluated when ranked in late positions. Due to the query's nature (retrieving facial images), users were interested in retrieving every possible true match. This is easily understood by considering the following scenario: performing a facial retrieving task on images stored in social networking databases and in adult's-content-tagged image databases in order to prevent violation of privacy. The NMRR measurement, due to its condition to consider every result retrieved after the K^{th} position as non-retrieved, often results in evaluating two different CBIR systems the same even if the correctly but late retrieved results are in very different positions.

Both the software used for the survey, as well as the results given by each participant, are available on-line[3]. Of course, given that the number of the participants is limited and the educational level is the same for all the individuals, further research and additional experiments are required in order to fully validate the observations arising from this case study.

5.4 Experiments on Benchmarking Databases

In order to encourage researchers in the field to use the proposed performance measure, MNRO has been implemented and is currently used in evaluating the retrieval results of the img(Rummager) system [19]. We have also implemented an application[4] which supports most of the standard measures used for evaluation in TREC, CLEF, and elsewhere, such as MAP, P(10), P(20) and BPref, as well as the ANMRR and the proposed ANMRO. Additional features include a batch mode and statistical significance testing (ST) of the results against a pre-selected baseline. STs tell us whether an observed effect, such as a difference between two means, or a correlation between two variables, could reasonably occur *just by chance* in selecting a random sample [51]. This application uses a bootstrap test, one-tailed [52], at significance levels 0.05, 0.01, and 0.001, against a baseline run. The results of the performance measures employed in the developed application correlate with the performance measure results of the TRECEval. TRECEval is the standard tool used by the TREC community for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results.

Finally, we present the experimental results in 3 known benchmarking databases for a large number of descriptors from the literature. We choose to calculate and evaluate the effectiveness of both global as well as local descriptors (bag-of-visual-words) in the Wang database, the UCID database and the ImageCLEF 2010 Wikipedia Database.

The Wang database is a subset of 1000 manually-selected images from the Corel stock photo database and forms 10 classes of 100 images each. The database is available on-line[5]. Although each category has its own semantic content, the visual content of images in one category could be very different. For this reason, the queries and ground-truths proposed by the MIRROR[40] image retrieval system are used. MIRROR separates the WANG database into

[3] http://www.ee.duth.gr/acsl/duthrobotics/index.html
[4] www.img-rummager.com
[5] http://wang.ist.psu.edu/docs/home.shtml

20 queries. The ground truth set is comprised of images from same category and with similar visual appearance. For example, the seventh set of the Wang database depicts horses. According to MIRROR, 'brown' horses forms a different query, with a different set of relevant images than the 'white' ones.

Descriptor	MAP	P(10)	P(20)	ANMRR	AMNRO
CEDD[47]	0.5891	0.6800	0.5500	0.2528	0.2773
FCTH[45]	0.5736	0.6450	0.5475	0.2737	0.2948
BTDH[46]	0.3503	0.4500	0.3600	0.5118	0.5496
C.CEDD[45]	0.5296	0.5900	0.5150	0.3064	0.3384
C.FCTH[45]	0.5222	0.6100	0.5175	0.3154	0.3467
JCD[53]	0.5880	0.6650	0.5500	0.2561	0.2783
SpCD[54]	0.4578	0.5450	0.4550	0.3841	0.4200
EHD[13]	0.3097	0.3650	0.3300	0.5264	0.5525
SCD[13]	0.2557	0.3400	0.2650	0.6117	0.6246
CLD[13]	0.4626	0.5150	0.4225	0.3927	0.4326
Color Histograms	0.3018	0.400	0.2925	0.5913	0.6160
Tamura Directionality[55]	0.2586	0.3100	0.2675	0.6154	0.6375
AutoCorrelograms[56]	0.3634	0.5050	0.4100	0.5011	0.5345
Top-Surf (10000)[57]	0.2526	0.3150	0.2750	0.6227	0.6429
Top-Surf (200000)[57]	0.1612	0.2350	0.1825	0.7654	0.7751

Table 5 Wang Database Results

Next, we performed experiments using the UCID database. The UCID database was created as a benchmark database for CBIR and image compression applications. UCID dataset is already widely being used for benchmarking CBIR algorithms [39][15][58][59]. This database currently consists of 1338 uncompressed TIFF images on a variety of topics including natural scenes and man-made objects, both indoors and outdoors. The UCID database is available for research[6]. All the UCID images were subjected to manual relevance assessments against 262 selected images, creating 262 ground truth image sets for performance evaluation.

Finally, we performed experiments on the ImageCLEF 2010 Wikipedia database. As mentioned in Section 5.2, this database consisting of 237,434 images and there are 70 test topics. From each topic we choose the first image as a query image. Query images are not part of the database.

In the same table, the results of a '*Text Only*' run were included in order to highlight that CBIR results are distinct from those of the text retrieval.

The results for these 3 databases are illustrated in Table 5, Table 6 and Table 7 respectively.

To show that the behavior of MNRO is not directly correlated with any of the 2 other image retrieval performance measures we performed the following experiment: We calculate how significant is the performance deviation between the descriptors in the Wang database. Indicative results are illustrated in Table 8.

[6] http://vision.cs.aston.ac.uk/datasets/UCID/ucid.html

Descriptor	MAP	P(10)	P(20)	ANMRR	AMNRO
CEDD	0.6748	0.2267	0.1237	0.2823	0.2224
FCTH	0.6723	0.2233	0.1208	0.2874	0.2315
BTDH	0.5353	0.1676	0.0912	0.4295	0.3957
C.CEDD	0.6584	0.2218	0.1221	0.2933	0.2284
C.FCTH	0.6487	0.2149	0.1191	0.3087	0.2402
JCD	0.6876	0.2290	0.1240	0.2683	0.2127
SpCD	0.5840	0.1859	0.1042	0.3791	0.3262
EHD	0.5326	0.1687	0.0931	0.4331	0.3852
SCD	0.4998	0.1565	0.0872	0.4667	0.4061
CLD	0.5361	0.1702	0.0947	0.4322	0.3694
Color Histograms	0.4443	0.1328	0.0718	0.5231	0.5051
Tamura Directionality	0.4411	0.1317	0.0748	0.5304	0.4978
AutoCorrelograms	0.5507	0.1721	0.0941	0.4139	0.3636
Top-Surf (10000)	0.4248	0.1344	0.0750	0.5462	0.5036
Top-Surf (200000)	0.3952	0.1229	0.0653	0.5788	0.5634

Table 6 UCID Database Results

Descriptor	MAP	P(10)	P(20)	ANMRR	AMNRO
Text Only	0.1291	0.3600	0.3300	0.7273	0.6974
FCTH	0.0062	0.0586	0.0507	0.9690	0.9205
SpCD	0.0056	0.0429	0.0421	0.9778	0.9293
CEDD	0.0055	0.0471	0.0450	0.9729	0.9255
C.CEDD	0.0047	0.0343	0.0321	0.9759	0.9271
C.FCTH	0.0038	0.0314	0.0314	0.9749	0.9265
EHD	0.0032	0.0271	0.0250	0.9827	0.9339
CLD	0.0030	0.0314	0.0307	0.9831	0.9342
Tamura Directionality	0.0011	0.0200	0.0171	0.9902	0.9418
Color Histograms	0.0007	0.0086	0.0050	0.9921	0.9431
SCD	0.0005	0.0157	0.0129	0.9929	0.9439

Table 7 ImageCLEF 2010 Wikipedia Database Results

	Descriptor(1)	Descriptor(2)	MAP	ANMRR	AMNRO
1	EHD	CLD	49.37%(***)	34.04%(**)	27.73%(**)
2	CH	CLD	53.25%(***)	50.56%(***)	42.41%(**)
3	FCTH	C.FCTH	9.85%(**)	15.23%(*)	17.61%(**)

Table 8 Performance Deviation Between Descriptors. Significance-tested with a bootstrap test, one-tailed, at significance levels 0.05(*), 0.01 (**), and 0.001 (***)

Based on these results, we observe that in Example 1, where we study the performance deviation between the Edge Histogram Descriptor (EHD) and the Color Layout Descriptor (CLD), MAP decides that the deviation is significant at level 0.001 while AMNRO and ANMRR, consider that the change is significant at level 0.01.

In Example 2, where we study the performance deviation between Color Histograms (CH) and the Color Layout Descriptor (CLD), AMNRO considers that the deviation is significant at level 0.01, while MAP and ANMRR, consider that the deviation is significant at level 0.001.

Finally, in Example 3, where we study the performance deviation between the Fuzzy Color and Texture Histogram (FCTH) and Compact Fuzzy Color and Texture Histogram (C.FCTH), AMNRO and ANMRR consider that the deviation is significant at level 0.01, while ANMRR, assumed that the deviation is significant at level 0.05.

In summary, we observe that AMNRO is not directly highly correlated with any of the 2 other image retrieval performance measures.

6 CONCLUSIONS AND FUTURE WORK

In this paper an overview of the most commonly used, single value performance measures for calculating the performance of retrieval systems was presented. The operating principles of Mean Average Precision and Average Normalized Modified Retrieval Rank were analyzed and their weaknesses were reported. Based on these weaknesses we proposed a new performance performance measure, called MNRO, which employs the sigmoid Gompertz function. The advantages of the new performance measure are demonstrated in several setups. In the first, artificially produced query trials and their evaluations were compared. A second experiment on a large database demonstrate the ability of the proposed performance measure to take into account the generality of the queries during the retrieval procedure. In the sequel, a subjective cross-evaluation of the image-retrieval results was performed by a group of 30 individuals. According to this experiment, in the vast majority of the cases the retrieval results of MNRO seem to be in agreement with what humans would select. Additionally, we present the experimental results produced by a large number of state of the art descriptors applied on three well-known benchmarking databases.

It is worth noting once again that, single value performance measures are used in order to compare different retrieval systems where most of the retrieval parameters, such as the database, ground truths, and scope are kept constant. In cases where it is preferable to evaluate the performance of a retrieval system using graphical representations, we suppose that the method proposed in [23] is the most comprehensive one, based on the fact that the generality parameter normalizes the precision vs recall graph.

The main criticism to MAP and ANMRR is that they are based on the assumption that retrieved data can be considered as either relevant or non-relevant to a user's information need. In the area of text retrieval, various measures have been developed which assign different levels of relevance to a given document [60–62]. In image retrieval, in order to evaluate systems with different levels of relevance the divergence function was introduced in [63]. This function evaluates the variance of a system ranking list to a user ranking list, which ranks the results depending on the different levels of relevance from the query. In these cases the user list is built based on the 'aboutness' [64, 65] of the images. An extension of our proposed method could emerge by incorporating a graded-relevance judgment property. A recently proposed method [29] gives MAP the capability to evaluate systems of different relevance grades. A relevant extension can be applied to both ANMRR and AMNRO.

The evolution of retrieval systems might lead to the development of systems which will require such performance measures.

Final, it is important to add to the MNRO the capability for evaluating systems with non complete judgments. Such types of databases often use BPref, which is highly correlated to MAP[66].

7 Acknowledgement

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. John R. Smith. Image retrieval evaluation. In *IEEE Workshop on Content-Based Access of Image and Video Libraries, 1998. Proceedings*, pages 112–113, 1998.
2. Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.
3. Henning Müller, Paul Clough, William R. Hersh, Thomas Deselaers, Thomas Martin Lehmann, and Antoine Geissbühler. Evaluation axes for medical image retrieval systems: the imageCLEF experience. In *ACM Multimedia*, pages 1014–1022, 2005.
4. Sharon McDonald, John Tait, and Ting-Sheng Lai. Evaluating a content based image retrieval system. In *SIGIR*, pages 232–240, 2001.
5. Joemon M. Jose, Jonathan Furner, and David J. Harper. Spatial querying for image retrieval: A user-oriented evaluation. In *SIGIR*, pages 232–240, 1998.
6. Farzin Mokhtarian, Sadegh Abbasi, and Josef Kittler. Efficient and robust retrieval by shape content through curvature scale space. *Series on Software Engineering and Knowledge Engineering*, 8:51–58, 1997.
7. Henning Muller, Wolfgang Muller, David Squire, Stephane Marchand-Maillet, and Thierry Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.
8. James Ze Wang, Jia Li, and Gio Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(9):947–963, 2001.
9. Gerald Schaefer and Michal Stich. Ucid: an uncompressed color image database. In *Storage and Retrieval Methods and Applications for Multimedia*, pages 472–480, 2004.
10. David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *CVPR (2)*, pages 2161–2168, 2006.
11. Mark J. Huiskes, Bart Thomee, and Michael S. Lew. New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. In *Multimedia Information Retrieval*, pages 527–536, 2010.
12. MPEG-7. *Subjective Evaluation of the MPEG-7 Retrieval Accuracy Measure (ANMRR).* ISO/WG11, Doc. M6029, 2000.
13. B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, and Akio Yamada. Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):703–715, 2001.
14. B. S. Manjunath, Philippe Salembier, and Thomas Sikora. *Introduction to MPEG-7: multimedia content description interface.* John Wiley & Sons Inc, 2002.
15. Thomas Deselaers, Daniel Keysers, and Hermann Ney. Features for image retrieval: an experimental comparison. *Inf. Retr.*, 11(2):77–107, 2008.

16. Avi Arampatzis, Konstantinos Zagoris, and Savvas A. Chatzichristofis. Dynamic two-stage image retrieval from large multimodal databases. In *ECIR*, pages 326–337, 2011.
17. Zhong Wu, Qifa Ke, Jian Sun, and Heung-Yeung Shum. Scalable face image retrieval with identity-based quantization and multireference reranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1991–2001, 2011.
18. Savvas A. Chatzichristofis and Yiannis S. Boutalis. Performance study of the most commonly used image retrieval evaluation methods. In *The Sixth IASTED International Conference on Advances in Computer Science and Engineering (ACSE)*, pages 27–32, 2010.
19. Savvas A. Chatzichristofis, Yiannis S. Boutalis, and Mathias Lux. Img(rummager): An interactive content based image retrieval system. In *SISAP*, pages 151–153, 2009.
20. Vijay V. Raghavan, Gwang S. Jung, and Peter Bollmann. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3):205–229, 1989.
21. Horst Eidenberger. Evaluation of content-based image descriptors by statistical methods. *Multimedia Tools Appl.*, 35(3):241–258, 2007.
22. Wessel Kraaij and Renée Pohlmann. Viewing stemming as recall enhancement. In *SIGIR*, pages 40–48, 1996.
23. Dionysius P. Huijsmans and Nicu Sebe. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):245–251, 2005.
24. Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
25. Mark Sanderson. Performance measures used in image information retrieval. In Henning Muller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors, *ImageCLEF*, volume 32 of *The Information Retrieval Series*, pages 81–94. Springer Berlin Heidelberg, 2010.
26. Adrian Popescu, Theodora Tsikrika, and Jana Kludas. Overview of the wikipedia retrieval task at imageCLEF 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
27. Henning Muller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors. *ImageCLEF - Experimental Evaluation in Visual Information Retrieval*. Springer, 2010.
28. Xiannong Meng. A comparative study of performance measures for information retrieval systems. In *ITNG*, pages 578–579, 2006.
29. Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. Extending average precision to graded relevance judgments. In *SIGIR*, pages 603–610, 2010.
30. Emine Yilmaz and Javed A. Aslam. Estimating average precision when judgments are incomplete. *Knowl. Inf. Syst.*, 16(2):173–211, 2008.
31. Stephen Robertson. A new interpretation of average precision. In *SIGIR*, pages 689–690, 2008.
32. Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. The maximum entropy method for analyzing retrieval measures. In *SIGIR*, pages 27–34, 2005.
33. Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *SIGIR*, pages 271–278, 2007.
34. Craig Macdonald, Iadh Ounis, and Ian Soboroff. Overview of the TREC 2009 blog track. In *The Eighteenth Text REtrieval Conference (TREC)*, 2009.
35. Mihai Lupu, Florina Piroi, Xiangji (Jimmy) Huang, Jianhan Zhu, and John Tait. Overview of the TREC 2009 chemical ir track. In *The Eighteenth Text REtrieval Conference*, 2009.
36. Alan F. Smeaton, Paul Over, and Aiden R. Doherty. Video shot boundary detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.
37. Jia Li and James Ze Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1075–1088, 2003.
38. Jens-Rainer Ohm. The MPEG-7 visual description framework - concepts, accuracy, and applications. In *CAIP*, pages 2–10, 2001.
39. Konstantinos Zagoris, Savvas A. Chatzichristofis, Nikos Papamarkos, and Yiannis S. Boutalis. img(anaktisi): A web content based image retrieval system. In *SISAP*, pages 154–155, 2009.

40. Ka-Man Wong, Kwok-Wai Cheung, and Lai-Man Po. MIRROR: an interactive content based image retrieval system. In *ISCAS (2)*, pages 1541–1544, 2005.
41. Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. The truth about corel - evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, CIVR '02, pages 38–49, London, UK, UK, 2002. Springer-Verlag.
42. Dionysius P. Huijsmans and Nicu Sebe. Extended performance graphs for cluster retrieval. In *CVPR (1)*, pages 26–33, 2001.
43. Klaas Bosteels and Etienne E. Kerre. Fuzzy audio similarity measures based on spectrum histograms and fluctuation patterns. pages 361–365, 2007.
44. Gerard Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
45. Savvas A. Chatzichristofis, Konstantinos Zagoris, Yiannis S. Boutalis, and Nikos Papamarkos. Accurate image retrieval based on compact composite descriptors and relevance feedback information. *IJPRAI*, 24(2):207–244, 2010.
46. Savvas A. Chatzichristofis and Yiannis S. Boutalis. Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. *Multimedia Tools Appl.*, 46(2-3):493–519, 2010.
47. Savvas A. Chatzichristofis and Yiannis S. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *ICVS*, pages 312–322, 2008.
48. Walid Magdy and Gareth J. F. Jones. Pres: a score metric for evaluating recall-oriented information retrieval applications. In *SIGIR*, pages 611–618, 2010.
49. Benjamin Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Phil. Trans. Roy. Soc. London*, 123:513–585, 1825.
50. Alberto dOnofrioa, Antonio Fasanob, and Bernardo Monechi. A generalization of gompertz law compatible with the gyllenbergwebb theory for tumour growth. *Mathematical Biosciences*, Article in Press, 2011.
51. David S. Moore, George P. McCabe, and Bruce Craig. *Introduction to the Practice of Statistics SPSS Manual*. WH Freeman, 2005.
52. Russell Davidson and James G. MacKinnon. Bootstrap tests: How many bootstraps? *Econometric Reviews*, 19(1):55–68, 2000.
53. Savvas A. Chatzichristofis, Avi Arampatzis, and Yiannis S. Boutalis. Investigating the behavior of compact composite descriptors in early fusion, late fusion, and distributed image retrieval. *Radioengineering*, 4:725–733, 2010.
54. Savvas A. Chatzichristofis, Yiannis S. Boutalis, and Mathias Lux. SpCD - spatial color distribution descriptor - a fuzzy rule based compact composite descriptor appropriate for hand drawn color sketches retrieval. In *ICAART (1)*, pages 58–63, 2010.
55. Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, 1978.
56. Jing Huang, Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. *US Patent 6,246,790*, 12:1–16, June 12 2001. US Patent 6,246,790.
57. Bart Thomee, Erwin M. Bakker, and Michael S. Lew. Top-surf: a visual words toolkit. In *ACM Multimedia*, pages 1473–1476, 2010.
58. Miguel Arevalillo-Herraez, Mario Zacares, Xaro Benavent, and Esther de Ves. A relevance feedback CBIR algorithm based on fuzzy sets. *Sig. Proc.: Image Comm.*, 23(7):490–504, 2008.
59. Daniele Borghesani, Costantino Grana, and Rita Cucchiara. Color features performance comparison for image retrieval. In *ICIAP*, pages 902–910, 2009.
60. Tetsuya Sakai and Noriko Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf. Retr.*, 11:447–470, October 2008.
61. Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October 2002.
62. W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Publishing Company, USA, 2009.
63. Jean Martinet, Shin'ichi Satoh, Yves Chiaramella, and Philippe Mulhem. Media objects for user-centered similarity matching. *Multimedia Tools Appl.*, 39(2):263–291, 2008.

64. Jean Martinet, Yves Chiaramella, and Philippe Mulhem. A relational vector space model using an advanced weighting scheme for image retrieval. *Inf. Process. Manage.*, 47(3):391–414, 2011.
65. Youngok Choi and Edie M. Rasmussen. Searching for images: The analysis of users' queries for image retrieval in american history. *JASIST*, 54(6):498–511, 2003.
66. Bilyana Taneva, Mouna Kacimi, and Gerhard Weikum. Gathering and ranking photos of named entities with high precision, high recall, and diversity. In *WSDM*, pages 431–440, 2010.