Entropy-Guided Assessment of Image Retrieval Systems: Advancing Grouped Precision as an Evaluation Measure for Relevant Retrievals

Tahar Gherbi[1], Ahmed Zeggari[1*], Zianou Ahmed Seghir[2] and Fella Hachouf[3]
[1]Math and Computer Sciences Dept. Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria
[2]Faculty ST, ICOSI Lab, University of Khenchela, Khenchela, Algeria
[3]Automatic and Robotic Lab, Electronic Dept. University of Mentouri Constantine, Constantine, Algeria
E-mail: tahar@univ-tebessa.dz, ahmed.zeggari@univ-tebessa.dz, zianou_ahmed_seghir@yahoo.fr, hachouf.fella@gmail.com
*Corresponding author

Keywords: information retrieval, CBIR, performance evaluation, precision, clustering, information theory, entropy

Received: February 6, 2023

The performance evaluation of Content Based Image Retrieval systems (CBIR), can be considered as a challenging and overriding problem even for human and expert users regarding the important numbers of CBIR systems proposed in the literature and applied to different image databases. The automatic measures widely used to assess CBIR systems are inspired from the general Text Retrieval (TR) domain such as precision and recall metrics. This paper proposes a new quantitative measure adapted to the CBIR particularity of relevant images grouping, which is based on the entropy of the returned relevant images. The proposed performance measure is easy to understand and to implement. A good discriminating power of the proposed measure is shown through a comparative study with the existing and well-known CBIR evaluation measures.

Povzetek: Članek predlaga novo kvantitativno mero za evalvacijo sistemov za iskanje slik na podlagi vsebine (CBIR), ki temelji na entropiji vrnjenih relevantnih slik.

1 Introduction

The aim of Content Based Image Retrieval (CBIR) systems [1] [2] is to rank the most similar images in the database given a user query and based on image content rather than textual annotations or metadata. A typical example of an image retrieval system, is when the CBIR system returns the relevant images from the database, in response to the image query of the user. Query by image content is an extremely active discipline; a large number of systems in different application areas are designed in the last twenty years. In [5], the authors report a tremendous growth in publications on this topic covering many disciplines such as medicine, botany, face recognition, fingerprint identification and place recognition. CBIR systems are based on automatic low-level image features extraction, such as color, gray shades and texture; not on a manual keywords annotation [3] and [4]. The evaluation of CBIR systems is based on benchmarking and performance metrics. The goal of a benchmark is to compare different systems on a set of test images database. An exhaustive survey on this topic can be found for example in [19], [21] and [20].

A main problem in the field of CBIR evaluation is the lack of a common performance measure that allows a quantitative and objective comparison of visual retrieval systems. The most used measures describe the number and/or the rank of relevant images within a returned list, Müller [8] and van Rijsbergen [16] present a good summary. Recent measures dedicated to CBIR system's evaluation have been proposed in the last few years. In [18] the authors proposed a new measure called: Mean Normalized Retrieval Order (MNRO) which uses the sigmoid Gompertz function to overcome the weaknesses of Mean Average Precision (MAP) and Average Normalized Modified Retrieval Rank (ANMRR) [17].

The density of returned relevant results is important and compatible with human vision evaluation. The common evaluation measures cannot illustrate the grouping propriety of the returned relevant images. In other words, the interrelation between relevant images was missed, which is important for a fast exploration of the result by a user visual inspection. For example, assuming a window size = 10, a system returning 100 images with 10 relevant images in one window is better than a system returning the same results with one relevant image by window. Furthermore, we extend the evaluation scale to achieve a better discriminating power in which two systems having a same precision value can be evaluated differently.

The rest of the paper is organized as follows. Section 2 provides an overview of the most used measures for information retrieval evaluation, section 3 describe the limitations of the standard measures especially for image retrieval. In section 3 we provides an outline on the proposed entropy based measure. Section 5 provides the experimen-

tal results and discussion. Finally, Section 6 draws the conclusions.

2 Measuring information retrieval: quantitative assessment

Quantitative evaluation measures in information retrieval (IR) are designed to fulfill specific criteria, including their correlation with user satisfaction, their ability to discriminate among retrieval results, and their ease of interpretation and implementation. These measures serve as valuable tools in assessing the performance and effectiveness of information retrieval systems.

The most widely used evaluation measures in IR are derived from the fundamental concepts of recall and precision. Recall represents the ability of a retrieval system to retrieve all relevant documents from a given data-set, while precision measures the proportion of retrieved documents that are truly relevant. These measures provide valuable insights into the accuracy and completeness of the retrieval results, enabling researchers and practitioners to assess and compare different systems or algorithms.

However, there are also alternative evaluation measures based on utility theory. These measures, as described in works such as [9, 10, 11], focus on measuring the worth or value of the retrieval output to the user. Utility-based measures take into consideration the utility or benefit that users derive from the retrieved documents, providing a different perspective on the quality of the retrieval system's output.

Utility-based measures are particularly useful in evaluating set-based retrieval output, as observed in tasks like the TREC filtering task [12]. By considering the worth of the retrieved documents to the user, these measures capture the relevance and usefulness of the retrieved set as a whole, rather than treating each document independently.

In a comprehensive evaluation scenario, an effective performance measure should adhere to the following criteria:

- Relevance of Retrieved Images: The measure should consider the number of relevant images returned by the system. It is essential that the retrieved images are indeed relevant to the user's query. This criterion ensures that the system accurately identifies and retrieves the desired content.
- Retrieval of Relevant Images: The measure should also take into account the size of the returned list. It is crucial that all relevant images are successfully retrieved by the system. A good performance measure should strive for high recall, aiming to retrieve as many relevant images as possible.
- Ranking of Relevant Images: The rank of the relevant images within the returned list is another important factor. The measure should prioritize placing the most relevant images at the top of the list. A higher-ranked position indicates a better performance, as it facilitates quick and efficient access to the most relevant content.
- Interrelations among Returned Relevant Images: The measure should consider the interrelations between the returned relevant images. Ideally, the relevant images should be grouped together rather than scattered throughout the list. This criterion ensures that the retrieval system provides coherent and meaningful results, enhancing the user's browsing experience.

By incorporating these criteria into the performance measure, researchers and practitioners can gain a comprehensive understanding of the system's effectiveness in information retrieval tasks. It allows for a holistic evaluation, considering relevance, retrieval completeness, ranking quality, and the overall organization of the retrieved content.

2.1 Mean average precision

Mean Average Precision (MAP) has been a popular evaluation metric in the field of Text Retrieval since its introduction in the Text Retrieval Conferences (TREC) starting from TREC 3 in 1994 [6]. Over the years, it has gained widespread adoption among researchers as a reliable measure for assessing the performance of their retrieval systems [7].

The MAP metric provides a comprehensive assessment by considering both precision and the ordering of relevant documents in the retrieval results. It calculates the average precision for each query and then takes the mean of these average precision values. The formula for MAP is as follows:

$$MAP = \frac{1}{R} \sum_{i=1}^{R} \frac{i}{r_i}$$

Here, R represents the total number of relevant documents in the entire collection for a specific information query. The term r_i denotes the ranking position of the ith relevant document in the retrieved list.

The MAP metric takes into account the ranking position of each relevant document. It assigns higher importance to relevant documents appearing at the top of the retrieved list. The formula calculates the precision at each position and then averages these precision values over all the relevant documents, providing a single numerical value that represents the overall performance of the retrieval system.

By utilizing MAP, researchers can evaluate the effectiveness of their retrieval systems by considering both the accuracy of the results (precision) and the completeness of the results (recall). It enables the comparison of different systems and the measurement of improvements made over time or across different experiments.

2.2 R-precision

The concept of R-precision provides a valuable insight into the performance of an information retrieval system by focusing on the precision achieved after retrieving a specific number, R, of relevant images for a given query. In other

words, *R*-precision measures the precision of the retrieval results up to a certain rank.

When *R* is equal to the total number of relevant images for the query, reaching an *R*-precision of 1.0 signifies an ideal scenario with perfect relevance ranking and perfect recall. It implies that all the relevant images in the collection have been retrieved within the top *R* positions, ensuring a complete and accurate representation of the query's intended information.

An *R*-precision value less than 1.0 indicates that not all the relevant images have been retrieved within the first *R* positions. This could be due to the presence of irrelevant or less relevant images in higher ranks, affecting the precision achieved. As the *R*-precision approaches 1.0, it signifies an improvement in the retrieval system's performance, as a larger proportion of the relevant images are appearing earlier in the retrieved list.

Evaluating the *R*-precision allows researchers and practitioners to assess the effectiveness and efficiency of their retrieval systems by examining how well the system ranks and retrieves relevant images at different points. It complements other evaluation measures like precision at different ranks, average precision, or mean average precision, providing a more granular understanding of the retrieval system's performance in the early stages of retrieval.

2.3 Precision and recall

The standard metrics used for evaluating the performance of information retrieval systems are precision and recall [15, 16]. Precision measures the proportion of relevant documents retrieved by the system out of all the documents that were returned. It provides an indication of the accuracy and relevance of the retrieval results. A high precision indicates that a large percentage of the retrieved documents are indeed relevant to the user's query.

On the other hand, recall measures the proportion of relevant documents that were retrieved out of all the relevant documents in the collection. It captures the system's ability to retrieve all relevant documents and reflects its completeness. A high recall suggests that a significant portion of the relevant documents has been successfully retrieved.

Precision and recall are complementary metrics that help assess different aspects of retrieval system performance. While precision emphasizes the quality of the retrieved results, recall emphasizes the system's ability to capture all relevant information. The balance between precision and recall depends on the specific requirements and goals of the information retrieval task.

By evaluating precision and recall, researchers and practitioners can gain insights into the effectiveness and efficiency of their information retrieval systems. These metrics allow for comparisons between different retrieval algorithms or system configurations, aiding in the optimization and enhancement of retrieval performance.

$$P = \frac{r(N)}{N}$$

In this context, the variable $r(N)$ denotes the count of relevant images retrieved, whereas N represents the size of the retrieved list. Precision is a straightforward evaluation measure that is often favored due to its ease of implementation. However, it does not take into account the specific rank positions of the relevant elements, making it less sensitive to their order in the retrieval results.

Recall is defined as the proportion of relevant documents retrieved from the database (*Rel*) out of all the relevant documents present.

$$R = \frac{r(N)}{Rel}$$

Ideally, a retrieval system should aim for high values for both precision (P) and recall (R) metrics. Rather than relying on individual measures of precision or recall, it is common to utilize a joint precision-recall (*PR*) graph to provide a comprehensive description of the system's performance [3]. The *PR* graph visually illustrates the trade-off between precision and recall at various thresholds or rankings.

However, one limitation of the *PR* graph is that its interpretation can be influenced by the number of relevant images associated with a particular query [8]. The shape and characteristics of the *PR* curve may vary depending on the specific query and the number of relevant images present. This means that comparing *PR* graphs across different queries or data-sets may not always provide a fair or meaningful comparison.

Despite this drawback, the *PR* graph remains a valuable tool for evaluating retrieval system performance. It allows researchers and practitioners to analyze the trade-off between precision and recall, make informed decisions about system parameters or algorithms, and understand the system's behavior at different retrieval thresholds. By considering the *PR* graph alongside other evaluation metrics, researchers can gain deeper insights into the strengths and weaknesses of their retrieval systems.

2.4 Recall-precision graph

A recall-precision graph is a graphical representation that illustrates the trade-off between recall and precision for a given information retrieval system or algorithm. Recall measures the completeness of the results returned by the system. It represents the proportion of relevant documents retrieved out of all the relevant documents in the collection. Higher recall indicates that more relevant documents are being retrieved. Precision, on the other hand, measures the accuracy of the retrieved results. It represents the proportion of relevant documents among all the documents retrieved. Higher precision indicates that a higher percentage of the retrieved documents are relevant. In a recall-precision graph, recall is typically plotted on the y-axis, while precision is plotted on the x-axis. The graph shows how the precision changes as the recall increases. The curve on the graph illustrates the relationship between recall and precision, and it can provide insights into the effectiveness of an information retrieval system. Ideally, a retrieval sys-

tem should achieve high precision and high recall simultaneously. However, in practice, there is often a trade-off between the two measures. The recall-precision graph helps to visualize this trade-off and assists in finding the optimal balance based on specific retrieval system requirements.

2.5 Entropy based measures

Entropy-based measures derived from the field of information theory play a significant role in the validation and evaluation of clustering algorithms. These measures provide valuable insights into the quality and effectiveness of clustering results. Among the various entropy-based measures, two popular ones commonly used are Entropy and Purity, proposed by Zhao and Karypis [13], and the V-measure proposed by Rosenberg and Hirschberg [14].

The concept of entropy, borrowed from information theory, provides a quantitative measure of the uncertainty or disorder within a cluster. It assesses how well the cluster's members are distributed across different classes or categories. Lower entropy indicates a higher degree of purity and cohesion within the cluster, suggesting that the members of the cluster predominantly belong to the same class.

Purity, on the other hand, measures the homogeneity of a cluster in terms of class labels. It evaluates how well the cluster assignments align with the true class labels of the data points. A high purity score signifies that the cluster contains predominantly instances from a single class, indicating a more accurate and reliable clustering result.

The V-measure combines both entropy and purity to provide a balanced evaluation metric for clustering. It captures the trade-off between homogeneity and completeness of a clustering solution. The V-measure is particularly useful when dealing with imbalanced data-sets, where some classes have a significantly larger number of instances than others.

By employing entropy-based measures such as Entropy, Purity, and the V-measure, researchers and practitioners can objectively assess the quality and coherence of clustering results. These measures help in comparing and selecting appropriate clustering algorithms, fine-tuning parameters, and optimizing the clustering process to obtain meaningful and accurate clusters.

$$Entropy = \sum_{i=1}^{K} \frac{k_i}{N} \left(-\frac{1}{\log C} \sum_{j=1}^{C} \frac{A_{ij}}{k_i} \log(\frac{A_{ij}}{k_i}) \right)$$

$$Purity = \sum_{i=1}^{K} \frac{1}{N} \max_{j}(A_{ij})$$

In the given context, the variables can be defined as follows: N represents the total number of data elements, C denotes the number of standard partitions, K signifies the total number of clusters, k_i refers to the size of cluster i, and A_{ij} indicates the count of elements in partition j that are assigned to cluster i.

The calculation of the $V - measure$ involves assessing the homogeneity and completeness of a clustering solution. These evaluations rely on entropy measures such as $H(C)$ and $H(K)$, as well as conditional entropy's including $H(C|K)$ and $H(K|C)$.

$$H(C) = -\sum_{j=1}^{C} \frac{\sum_{i=1}^{K} A_{ij}}{N} \log \frac{\sum_{i=1}^{K} A_{ij}}{N}$$

$$H(K) = -\sum_{i=1}^{K} \frac{\sum_{j=1}^{C} A_{ij}}{N} \log \frac{\sum_{j=1}^{C} A_{ij}}{N}$$

$$H(C|K) = -\sum_{i=1}^{K} \sum_{j=1}^{C} \frac{A_{ij}}{N} \log \frac{A_{ij}}{\sum_{j=1}^{C} A_{ij}}$$

$$H(K|C) = -\sum_{i=1}^{K} \sum_{j=1}^{\downarrow} \frac{A_{ij}}{N} \log \frac{A_{ij}}{\sum_{i=1}^{K} A_{ij}}$$

3 Shortcomings of conventional quantitative metrics in evaluation

The standard quantitative metrics fail to consider certain crucial factors that are vital for a comprehensive quantitative evaluation of content-based image retrieval systems. Firstly, they overlook the significance of high density of relevant results, where relevant images are clustered together within a small or large collection area in the retrieved window. This characteristic is not adequately captured by existing evaluation metrics, which can be attributed to their origins as general information retrieval (IR) measures.

Secondly, the discriminating power of a quantitative evaluation metric is often overlooked. This raises an important question: if two retrieval results have the same precision value, does it imply that they are similar? In other words, can we evaluate their corresponding systems as identical?

Considering these points from our perspective, it becomes evident that the existing evaluation metrics might not fully address the nuances and complexities of content-based image retrieval. There is a need for more refined metrics that take into account factors such as the clustering of relevant results and the ability to differentiate between retrieval outcomes with similar precision values. By developing and incorporating such metrics, we can improve the accuracy and effectiveness of quantitative evaluations in content-based image retrieval systems.

Table 1 shows the level of respect to the up cited proprieties by different measures used in this study.

In the following subsections, we discussed in detail these points which must be verified by our proposed CBIR evaluation measure.

3.1 Relevant results density

Verification of the pertinent results in the case of image retrieval, is much different from that of the general infor-

Table 1: A summary table of the measures used in this study in response to the proprieties of number, ranking and grouping.

Results	Number	Ranking	Grouping
Precision(P)	High	No	No
MAP	Medium	Medium	No
R-precision R(P)	No	High	No
RBP	Medium	Medium	Medium

mation retrieval, from which common evaluation measures are inspired. In the case of textual results search, the verification of pertinent results among the returned list must be done on a sequential manner, from the first result to the last one [22]. However, the visual verification of pertinent images is by nature very fast, and guided by the location and the grouping of relevant images. An evaluation process starts with an inherent transformation of the returned results to a binary list containing *relevant* and *irrelevant* items. Figure 1 display some user query results (sad and happy emojis) [22]. Even if the results in the first results are more precise (27,77%) than those in second one (precision = 22,22%), the presentation of the returns in the first results is difficult for the user to evaluate and verify. However, when the findings are gathered together, even with a lower precision rate, the results are considerably better for user evaluation. Additionally, it should be noted that, in contrast to the first results, the relevant images in the second results are located near the bottom of the 2D list. The results in this example are binary (either a sad or happy emoji), however in the real situations, the scenario is far more complex and has more than two potential outcomes. Another problem, is what we called situation search. Asking a system to return only one image from a database containing many relevant images raise almost to a full precision. In the next two subsections we study the effects of relevant images on the evaluation process when two systems have a same precision rate.

3.2 Comparing results having a same full precision

An ideal CBIR system provides a perfect image retrieval results, in which each image query returns a list of relevant results with no prior knowledge about its size. Its size varies from an input image to another input image. Therefore, the returned list = the relevant list of a given query in the database. Let $P(R, N)$ the precision of a retrieval result, where R represents the number of relevant images, and N represents the size of the returned list. We distinguish two evaluation cases regarding the number of relevant images contained in the database:

– The effectiveness of a system when the database contains a few relevant images, In this situation of full precision ($P = 1$), the amount of getting precision increase when R decrease. A minor error of retrieving R affect the systems precision not equitably. Hence, we define relevant error R_ERR, as a minimum precision mistake taken by a given system reducing R as:

$$R\_ERR = \frac{R-1}{N}$$

– The effectiveness of a system when the database contains many relevant images. When we have a large number of relevant images in a database, it is very challenging to have a full precision from a large returned list than from a small one. Hence, we define retrieval error N_ERR, as a maximum no zero precision mistake taken by a system as:

$$N\_ERR = \frac{1}{N}$$

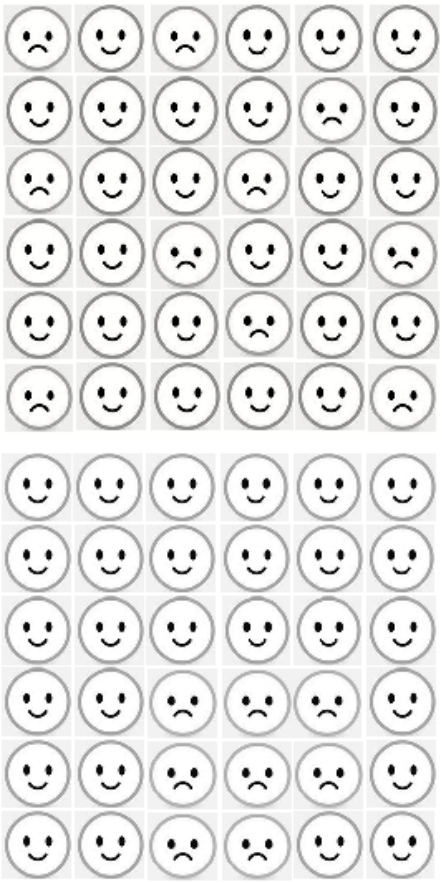


Figure 1: Example of two returned lists: dispersed results and grouped results.

The best retrieval situation is the system that returns one and only one exact image. This system has the highest risk, in which precision has a binary value (0 or 1). The next best system is a system in which the size of the final returned list is higher. In that case, the risk to obtain no relevant image is higher than in the case of a smaller returned list. We can define a precision error P_ERR as follows:

$$P\_ERR = \min(R\_ERR, N\_ERR).$$

P_ERR represents the minimum of the two errors as shown in figure 2 in which the minimum errors intersection is depicted.

3.3 Comparing results having a same precision ($P < 1$)

Two results having a same precision value $P_i(R_i, N_i) = P_j(R_j, N_j)$ can be evaluated differently when $R_i \neq R_j$. System i is better than system j when $R_i < R_j$, because the minimum size of returned list needed by system i to return the same number of relevant images R_j is: $N_i + R_j - R_i < N_j$. In that case, the new precision becomes: $P'_i = \frac{R_j}{N_i + R_j - R_i} > P_j$.

4 An entropy based development for visual retrieval systems

It is important and useful for a user to see the images that he needs, arranged together in the same part on the returned list. The main idea of a new measure project is to evaluate retrieval systems regarding a degree in which relevant images are grouped, which is very practical for a visual retrieval perspective.

The proposed Entropy Grouped Relevant images measure (EGR) is initially presented as follows:

$$EGR = -\sum_{i=1}^{|K|}\sum_{j=1}^{|C|}\frac{A_{ij}}{N \times c_j}\log\frac{A_{ij}}{N \times c_j}$$

Where: N represent the number of returned images, R is the number of relevant images in the returned list. K and C are the sets of the detected clusters and the standard partitions respectively. A_{ij} is Number of elements that are members of cluster i and partition j of the same class. Figure 3 show an example of returned results composed of clusters (a) and partitions (b).

5 Experimentation's

In order to evaluate the proposed measure, we compare it with other precision-based measures, including standard precision P, MAP, R-precision and RBP [23] measures. The comparison process is built around two tests: comparison based on a fixed size of a returned list, and comparison based on different sizes.

5.1 Comparison based on a fixed size of a returned list

In this stage, we conducted a comparison study in a situation when the system returns 12 images as returned linear list. For example when $R = 10$, the possible results in a linear list are: (10), (9,1), (8,2), (7,3), (8,1,1), (6,4), (5,5), (7,2,1), (6,3,1), (6,2,2), (5,3,2), (4,4,2), (4,3,3). They have a same precision value $P = 83,3$. The *EGR* measure evaluates differently these results according to the spatial density of the relevant items arrangement.

5.1.1 Best ranks on a fixed returned list

As can be seen in table 2, the results are ordered and ranked according to the different measures used in this comparative study. The top five results are well ranked by EGR and RBP measures, the best five results are highlighted in bold. Theirs ranks correspond well to the user ranking and to the real position of these results. The other measures ranked theme on the first three ranks. The discriminating power of the proposed measure and RBP measure appears in the ranking of the five best results on the five best ranks. Whereas, precision (P) for example ranks 52 results on five best places (which correspond to 55% of all results).

Table 2: Some selected results of the five best ranks according to five evaluation metrics when the returned list size N=12.

Rks	P	MAP	R(P)	RBP	EGR
1^{st}	**12**	**12-11-10**	**12-11-10**	**12**	**12**
2^{nd}	**11-(10,1)**	**(10,1)**	**(10,1)**	**11**	**11**
3^{rd}	**10-(9,1)**	**(9,1)**	**(9,1)**-(8,2)	**(10,1)**	**(10,1)**
4^{th}	9-(8,1)-(7,2)	(8,1)	(8,1)-(7,2)	**10**	**10**
5^{th}	8-(7,1)-(6,2)	(7,1)	(7,1)-(6,2)	**(9,1)**	**(9,1)**

5.1.2 Worst ranks on a fixed returned list

The superiority of the proposed entropy based measure, over the other measures to interprets the worst results ranking appears on table 3. The verified worst results appears individually on each rank on EGR measure. Whereas, it appears with other results in the case of precision measure (P). The other measures (ie, MAP, R and RBP measures) cannot capture this results as the worst places.

Table 3: Some chosen results of the five last ranks according to five evaluation metrics when the returned list size N=12.

Rks	P	MAP	R(P)	RBP	EGR
91^{th}	5-(4,1)-(3,2)-(3,1,1)	**(1,1,1)**	**(1,1,1)**-(2,1,1,1,1)	5	**(2,1)**
92^{th}	4-(3,1)-(2,2)-(2,1,1)	(2,1,1,1,1,1)	(2,1,1,1,1,1)	(1,2,1)	**(1,1,1)**
93^{th}	3-**(2,1)-(1,1,1)**	(1,1,1,1)	(1,1,1,1)	**(2,1)**	**2**
94^{th}	**2-(1,1)**	(1,1,1,1,1)	(1,1,1,1,1)	**(1,1)**	**(1,1)**
95^{th}	**1**	(1,1,1,1,1,1)	(1,1,1,1,1,1)	**1**	**1**

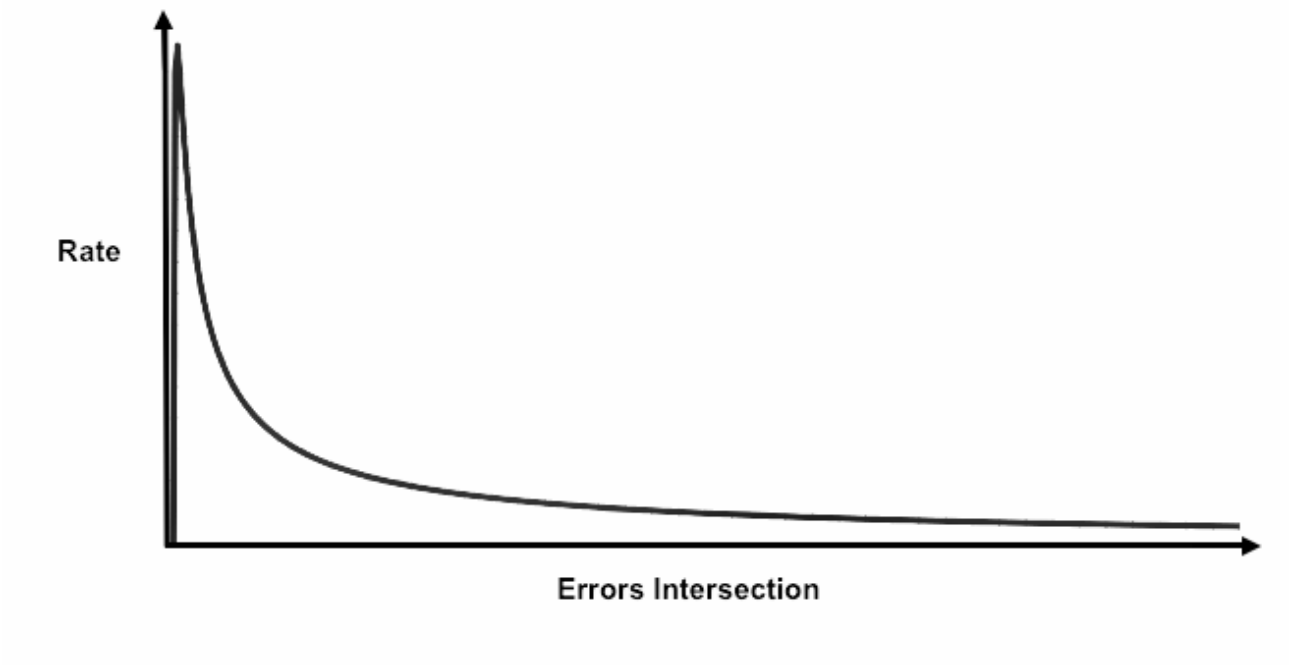


Figure 2: *P_ERROR* according to minimum error rates, *P_ERROR* is the same as *N_ERROR* except when N=1

5.2 Comparison based on different sizes of a returned list

The first comparison is built around the effectiveness of the proposed measure to evaluate systems having different sizes of returned images which are all relevant.

As can be seen from figure 4, the best system is when $R = N = 1$; *i.e.*, target search. The next best systems are ordered according to the greatest value of their returned lists. Such results corresponds well to $P - ERROR$ depicted in figure 2.
Table 4 summarize the five evaluation measures when $R = N$.

5.2.1 Some returned images are relevant ($R \leq N$)

The first remark can be seen from table 4 is that a results are ordered according to *EGR* values, they are well corresponds to the human order than the other measures.

There is an attempts to compare this results even that they have different natures (different sizes and different numbers of relevant images returned). *EGR* values are very closes when theirs corresponding results are perceptually very closes. Inversely, they are very different when theirs corresponding results are very distinct.

Table 4: The best full precision results arranged by *EGR* measure.

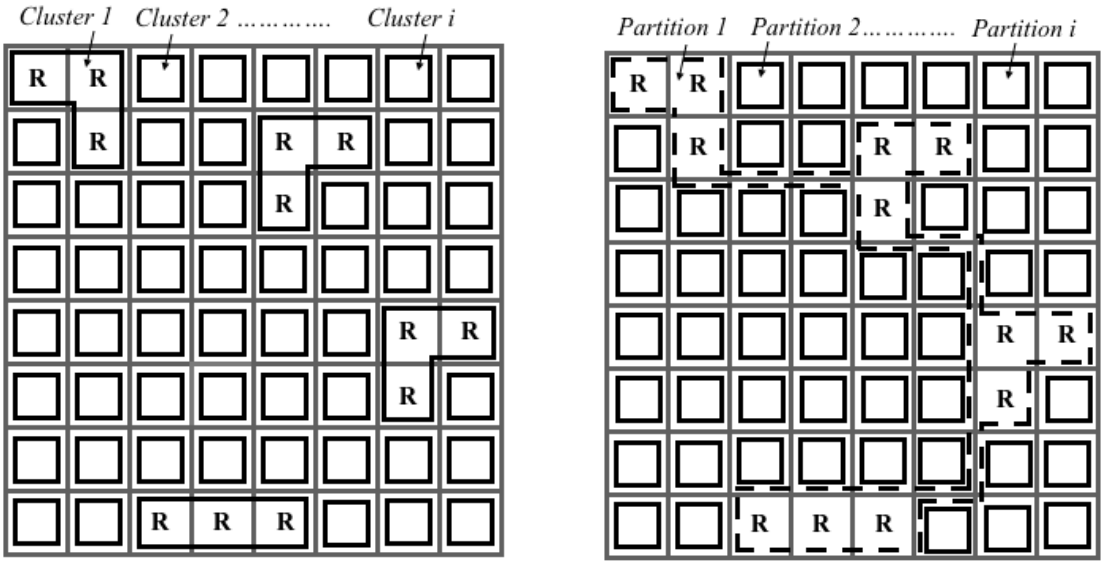
Retrievals	EGR	P (%)	MAP (%)	R(P) (%)
1,1	0	100	100	100
19,19	0,067	100	100	100
18,18	0,069	100	100	100
15,15	0,078	100	100	100
11,11	0,094	100	100	100
5,5	0,139	100	100	100

6 Conclusions

We have proposed a new evaluation measure to assess image retrieval systems. The proposed metric is compatible and conforms with human vision evaluation. In addition to the number and the rank of the relevant images on the returned list, the proposed measure can capture and enhance the presence of relevant images in a close area of the returned list. Based on entropy of pertinent images grouping, the proposed measure presents a high discriminating power against several retrieval cases, in which the actual measures evaluate them as equivalent. This allows us to use the proposed CBIR evaluator as a scale rather than an evaluation metric. Further investigations and experiments should be conducted, encompassing diverse situations and scenarios, to establish a robust and reliable performance measure for the proposed metric in the field of image retrieval. Additionally, its applicability in other domains, such as image quality assessment and data clustering.

References

[1] S. Selvakanmani, Ashreetha B, G. Naga Rama Devi, Shubhrojit Misra, Jayavadivel R, Suresh Babu Perli, Deep learning approach to solve image retrieval issues



(a) detected clusters (56 clusters) (b) standard partitions (53 partitions)

Figure 3: Example of two returned results and theirs corresponding clusters and partitions.

associated with IOT sensors, Measurement: Sensors, Volume 24, 2022, 100458, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2022.100458

[2] K. Wangi and A. Makandar, Autoencoder for Image Retrieval System using Deep Learning Technique with Tensorflow and Kears, 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 2023, pp. 1-5, https://doi.org/10.1109/ICICACS57338.2023.10099675

[3] Yong Rui and Thomas S. Huang (1999) Image retrieval: Current techniques, promising directions and open issues, *Journal of Visual Communication and Image Representation*, v.10, pp. 39–62. https://doi.org/10.1006/jvci.1999.0413

[4] Smeulders, Arnold W. M. and Worring, Marcel and Santini, Simone and Gupta, Amarnath and Jain, Ramesh (2000) Content-Based Image Retrieval at the End of the Early Years, *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, v.22(12), pp. 1349–1380. https://doi.org/10.1109/34.895972

[5] Datta, Ritendra and Joshi, Dhiraj and Li, Jia and Wang, James Z. (2008), Image Retrieval: Ideas, Influences, and Trends of the New Age, *ACM Comput. Surv.*, ACM Publisher, v.40(2), pp. 5:1–5:60. https://doi.org/10.1145/1348246.1348248

[6] url = http://trec.nist.gov/.

[7] Wu, Shengli and McClean, Sally (2006), Information Retrieval Evaluation with Relevance Judgment, *Flexible and Efficient Information Handling*, Springer Berlin Heidelberg, pp.86–93. https://doi.org/10.1007/11788911-7

[8] Müller, Henning and Müller, Wolfgang and Squire, David McG. and Marchand-Maillet, Stephane and Pun, Thierry (2001) Performance Evaluation in Content-based Image Retrieval: Overview and Proposals, *Pattern Recogn. Lett.* Elsevier Science Inc., v.22(5), pp.593–601. https://doi.org/10.1016/s0167-8655(00)00118-5

[9] Cooper, William S. (1973) On selecting a measure of retrieval effectiveness, *Journal of the American Society for Information Science*, v.24(2), pp. 87-100. https://doi.org/10.1002/asi.4630240204

[10] Lewis, David D. (1995) Evaluating and Optimizing Autonomous Text Classification Systems, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Seattle, Washington, USA, pp.246–254. https://doi.org/10.1145/215206.215366

[11] Buckley, Chris and Voorhees, Ellen M. (2000) Evaluating Evaluation Measure Stability, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*, ACM, Athens, Greece, pp. 33–40. https://doi.org/10.1145/345508.345543

[12] David D. Lewis (1995) The TREC-4 Filtering Track, *TREC*, National Institute of Standards and Technology (NIST), Special Publication 500-236.

[13] Zhao, Ying (2005) Criterion Functions for Document Clustering, *phd thesis* University of Minnesota USA.

[14] Rosenberg, Andrew and Hirschberg, Julia (2007) V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp.410–420.

[15] Salton, G. (1971) The SMART Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall, Inc., Upper Saddle River, NJ, USA. https://doi.org/10.1109/tpc.1972.6591971

[16] C. J. van Rijsbergen (1979) Information Retrieval, Butterworth. https://doi.org/10.1002/asi.4630300621

[17] MPEG-7 (2000) Subjective evaluation of the MPEG-7 retrieval accuracy measure (ANMRR). *ISO/WG11, Doc. M6029*.

[18] Savvas A. Chatzichristofis and Chryssanthi Iakovidou and Yiannis S. Boutalis and Elli Angelopoulou (2014) Mean Normalized Retrieval Order (MNRO): a new content-based image retrieval performance measure, *Multimedia Tools Appl.*, v.70(3), pp.1767–1798. https://doi.org/10.1007/s11042-012-1192-z

[19] Harald Kosch and Paul Maier (2010) Content-Based Image Retrieval Systems - Reviewing and Benchmarking, *JDIM*, v.8(1), pp.54–64.

[20] Henning Müller, Antoine Geissbuhler, Stephane Marchand Maillet, Paul Clough (2004) Benchmarking image retrieval applications, *Workshop on Visual Information Systems*, pp.334–337.

[21] Ezekiel Mensah Martey, Hang Lei, Xiaoyu Li, Obed Appiah (2021), Effective Image Representation Using Double Colour Histograms For Content-Based Image Retrieval, *informatica*, vol.45 No.7

[22] T. Gherbi, A. Zeggari and Z. Ahmed Seghir (2023), A global precision view for information retrieval evaluation adapted to image retrieval systems, *ICAECE'2023, Tebessa*, AIJR Publisher,

[23] A. Moffat and J. Zobel (2008). Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst., 27:2:1–2:27, 2008. doi/10.1145/1416950.1416952