Performance Evaluation of Relevance Feedback for Image Retrieval by "Real-World" Multi-Tagged Image Datasets

*Roberto Tronci *, AmILAB - Laboratorio Intelligenza d'Ambiente Sardegna Ricerche, ITALY*

Luca Piras, DIEE - University of Cagliari, ITALY

Giorgio Giacinto, DIEE - University of Cagliari, ITALY

ABSTRACT

Anyone who has ever tried to describe a picture in words is well aware that it is not an easy task to find a word, a concept or a category that characterizes it completely. Most of the images in real life represent more than a concept, and it is therefore natural that images available to users over the Internet (e.g., FLICKR) are associated with multiple tags, where with the term 'tag' we refer to a concept represented in the image. The purpose of this paper is to evaluate the performances of relevance feedback techniques in content-based image retrieval scenarios with multi-tag datasets, as typically performances are assessed on single-tag dataset. Thus, we show how relevance feedback mechanisms are able to adapt the search to user's needs either in the case an image is used as an example for retrieving images each bearing different concepts, or the sample image is used to retrieve images containing the same set of concepts. In this paper we also propose two novel performance measures aimed at comparing the accuracy of retrieval results when an image is used as a prototype for a number of different concepts.

Keywords: *content based image retrieval, relevance feedback, correlation measures, performance evaluation, multi-tagged datasets*.

INTRODUCTION

The retrieval of images from multimedia collections is one of the most active topics in computer science nowadays. In many cases, e.g., a raw collection of images from a photo repository, the only information that can be extracted is the related visual content. In these cases the retrieval task must rely only on low-level features of the images and, as a consequence, the task becomes more difficult due to the semantic gap between the visual content (i.e., the data used for the retrieval process) and the semantic concepts (i.e., the goal of the retrieval process) (Smeulders et al., 2000; Lew et al., 2006; Datta et al., 2008). All the techniques that are based on the visual content of the images fall in the Content Based Image Retrieval (CBIR) field.

It has been shown that the effectiveness of a CBIR technique strongly depends on the choice of the set of visual features, and on the choice of the metric used to model the users' perception of image similarity. Unfortunately, it is a very hard task to assess which are the "best" visual features for a given retrieval task, and which is a suitable similarity metric. Consequently, the set of retrieved images often fits the users needs only partly. To overcome this problem, the use of Relevance

Feedback has been widely studied as a tool to allow users refining the results by submitting a feedback on the images' relevance (Huang et al., 2008; Zhou & Huang, 2003; Rui & Huang, 2001; Giacinto, 2007).

Relevance Feedback (RF) is a mechanism that involves directly the user by allowing her to refine the retrieval results by marking the retrieved images as relevant or non-relevant to the visual query. Then, this feedback is exploited to "adjust" the retrieval mechanism, and it is used to propose to the user a new set of images that is deemed to be relevant according to the given feedback. This "adjustment" can be made in different ways, but basically they can be subdivided into two methodologies: one that is based on some transformations of the visual feature space, and the other that is based on the modification of the similarity metric, both aiming to attain higher similarity between relevant images, and lower similarity between relevant and non-relevant images.

Relevance Feedback approaches are usually based on the formulation of a two-class problem where the retrieved images are labelled as being either relevant or non relevant to the query. The behavior of RF techniques following this problem formulation is usually evaluated by means of images' datasets where each image is associated with one tag (i.e. the label associated with a visual concept), even if the images typically contain more than one visual concept.

It is the opinion of the authors that the assessment of the capabilities of RF techniques carried out by this kind of datasets is somehow limited. This limitation is related to the fact that only a single tag/concept is associated with each image in the dataset. Commonly, this "limitation" is accepted for two reasons: first, the single tag is associated with the most significant visual concept of the image as it is easier to find; second, it is easier, during the testing phase, to artificially simulate the behavior of users that submit the feedback with respect to only one concept. Despite the fact that these reasons allow to set up a simple evaluation scenario, they do not allow to fully evaluate the potential behind RF techniques because an image is typically associated with a number of concepts, and because this modality of simulated feedback is not a good model of the behavior of a real user.

The problem of multi-tagged datasets has been faced in different fields (Tsoumakas et al., 2010), such as protein function classification, music categorization, and also semantic scene classification (Boutell et al., 2004), but unfortunately it remains an aspect almost not tackled in the field of Relevance Feedback for Content Based Image Retrieval. Probably, the main reason is that the simulation of a real user's feedback in a real case scenario (i.e. when different visual concepts are associated with a single image) is quite difficult. Following the query-by-example paradigm, the user submits to the CBIR system an image as a query, and the system retrieves the images that are most similar to the query from the visual point of view. When the user is asked to mark the images as being relevant or not, then a real user can be interested in retrieving images related to more than just one of the visual concepts represented in the query image. Thus, different retrieval tasks can be performed starting from a given query image, as many as the possible combinations of visual concepts contained in the image itself. This kind of behavior is quite difficult to simulate unless you have a toy dataset, or a large number of real users agree to perform a live experimentation.

In this paper we propose a "new" modus operandi for evaluating a Relevance Feedback methodology in multi-tagged datasets. To this aim we propose different ways to simulate some of the possible behaviors of real users in different scenarios, i.e., the logic employed by the user to mark the images as being relevant or not. In our opinion these scenarios can be used to attain a more thorough assessment of RF techniques. Moreover we propose some novel measures of concept correlation, aimed at assessing if the result obtained in the experiments of RF techniques in different scenarios is reliable, or if it is biased by one of the single concept. In fact, while in the case of a single-tagged dataset the retrieval process is clearly driven by just one concept, in the case of a multi-tagged dataset it can be of interest to assess if the search for multiple concepts is actually driven by a few of them, or if these multiple concepts actually represent a higher level concept.

The paper is organized as follows. Firstly, we propose our view/ideas on evaluating a relevance feedback methodology in multi-tagged datasets. Secondly, we briefly review the RF techniques that

are used in the experimental phase. After these sections, we give the details of the experimental phase, and finally the conclusions are outlined.

EVALUATING A RELEVANCE FEEDBACK METHODOLOGY IN REAL WORLD DATASETS

In the introduction, we have briefly described how the Relevance Feedback interaction works. This interaction is fully detailed in the schema proposed in Figure 1:

1. An user submits a query to the system by a sample image
2. The CBIR system extracts the visual features from the query
3. The CBIR system computes the similarities of the query with the whole image database
4. The CBIR system proposes to the user the first k images in terms of similarity
5. The user submits its feedback by labelling each of the k images as being relevant or non-relevant
6. The RF system exploits the feedback to re-compute the similarities, and proposes to the user other k images [The steps 5-6 can be repeated with no limitations]

Figure 1: The general schema of a CBIR system with RF.

Typically, experimental verifications are carried out by automatic simulation of users' behavior to ensure test repeatability. As we have already stated in the introduction, the "usual scenario" is the one of using single tagged datasets (i.e. one concept/tag per image), thus user's feedback is simulated by exploiting the tags. In this case, the "simulated user" provides a unique feedback per image (i.e., by using the single tag associated with that image) even if a number of semantic concepts are present in the image. These kinds of datasets/experimentations are often not "realistic" as they just focus on one of the concepts represented in each image. Thus, the use of real world (i.e. multi-tagged) datasets/experimentations should be used to assess the performance of the proposed methodologies in real world scenarios. To this aim, the use of such a kind of datasets is progressively increasing in the field of Image Retrieval. In the introduction we have also reported some papers that proposed the use of multi-tagged dataset, but none of them have faced the use relevance feedback in content-based image retrieval.
In the following, we will propose a "modus operandi" on how the evaluation of a relevance feedback methodology should be carried out according with the authors' view.

New testing scenarios for multi-tagged image datasets

In this section we will describe some new testing scenarios to be used for multi-tagged image datasets. It is worth noting that the use of multi-tagged image datasets have some drawbacks, especially in the case of the assessment of relevance feedback methodologies, as they complicate the simulation of a user's feedback with respect to the "usual scenario". As already stated, the "usual scenario" allows an easy automatically "simulation" of the feedback given by user, as the query image is associated with just one concept, and the simulated user marks as relevant all the images that belong to the same concept of the query, and marks as non-relevant all the other ones. On the other hand, in a real-world scenario, each image is associated with more than one concept, thus some questions arise: *can we simulate the behavior of a user giving her feedback? If yes, how? If no, are there some ways to partially simulate it?* The "modus operandi" that we propose aims to answer to these questions.
Let us take into account a dataset where each image is associated with different *concepts* $L = \{l_1, l_2, ..., l_n\}$. Let us take an image q from the dataset as a query, and let us assume that concepts $\{l_1, l_2, ..., l_n\}$ are associated with q, where $k < n$. Users who starts the search using q as a query can follow one of the four scenarios described below:

- **Single concept scenario:** this scenario is related to users that are interested to just one of the k concepts $\{l_1, l_2, ..., l_k\}$. This scenario is quite similar to the usual scenario, with the difference that each image belongs to different concepts, thus different sets of images can be relevant to the query depending on the selected concept. In the literature of multi-label classification this scenario is referred to as Binary Relevance learning.
- **Multiple concepts AND scenario:** in this scenario users are interested in images that exhibits an exclusive combination of the concepts related to the query (pairs, terns, etc.) $\{l_1 \wedge l_2, ..., l_{k-1} \wedge l_k, l_1 \wedge l_2 \wedge l_3, ...\}$. In this scenario the search for similar images is driven by the combination of concepts. It is easy to see that the search for multiple concepts is analogous to the search of a single higher level concept made up of the combination of concepts l_c. This scenario is usually referred to as Label Powerset in the literature of multi-label classification.
- **Multiple concepts OR scenario:** in this scenario users are interested in images that exhibits a non-exclusive combination of the concepts related to the query $\{l_1 \vee l_2, l_1 \vee l_3, ..., l_2 \vee l_3 \vee l_k, ...\}$. Thus, an image is relevant if it exhibits at least one of the concepts the user is looking for. E.g., in the case of $\{l_1 \vee l_2 \vee l_k\}$, at each interaction the images that are relevant to the user's query are those who belong either to one of the concepts $\{l_1, l_2, l_k\}$ or to a combination of them.
- **Multiple concepts AND-OR scenario:** this is the combination of the previous two scenarios. In this case the target of the search is a combination of concepts connected by AND and OR statements: e.g., a possible target is $\{l_1 \vee (l_2 \wedge l_3)\}$, and the related relevant images are those that have l_1 or l_2, l_3 or l_1, l_2, l_3 as concepts.

These four scenarios cover all possible behaviors of a real user. All these scenarios assume that the user holds the same behavior for a given query, and for each feedback interaction. In a real case, a user can also switch between scenarios in different interactions, e.g. she starts using the pair $\{l_1 \wedge l_2\}$, while afterward she starts marking as relevant also the images that exhibit just the l_1 concept.
The single concept scenario allows testing the ability of a relevance feedback methodology not only to adapt the search to the feedback during the marking interaction, but also to verify the capability of adaptation to different target concepts even if the query image is the same.
The multiple concepts AND scenario tests the ability of a relevance feedback methodology to refine the search exploiting the feedback when the target is a combination of visual concepts. From a formal point of view, the AND of different concepts produces a new concept l_c (i.e. is formally equivalent to the previous one).
The multiple concepts OR scenario and *multiple concepts AND-OR* scenario verify the ability of relevance feedback techniques to refine the search when the target is a "complex" combination of visual concepts. It is worth noting that the *multiple concepts AND-OR scenario* is nothing but the *multiple concept OR scenario* with more complex single concepts.
Thus, in our view, the experimental evaluation phase of a novel relevance feedback technique should take into account the above scenarios.

"Post-correlation" measures in multiple concept query scenarios

The use of the scenarios proposed in the previous section can offer different point of views for the analysis of relevance feedback techniques. In fact, the use of multi-tagged image datasets and the use of different "simulations" of user logics, allow a researcher to extensively prove the adaptation capabilities of the relevance feedback techniques that has been devised.
As partially mentioned in the previous sections, RF approaches work in this way: the system presents a limited number of images n to the user; then the user marks these images as being relevant or not, and this feedback is exploited by RF to recalibrate the retrieval results and propose "new" images to the user (the user can iterate the process as many times as she wants). Thus, in the scenarios proposed for a multi-tagged dataset, the feedback is provided by means of a combination

of concepts: in this case, *is it the operation of recalibrating the results driven by the combination of concepts or by a subset of them from a formal point of view?*
Moreover, when we are looking for a combination of concepts, is the feedback given for a combination of concepts better than the feedback given for the individual concepts in terms of relevant images that are found after the recalibration step? These questions are not related to the evaluation of the performance of relevance feedback techniques, but they are useful to analyze in a deeper way the behavior of relevance feedback techniques with respect to a given image collection, to assess the presence of biases in the evaluation process.
To provide an answer to the previous questions, we propose to use the following concept *correlation measures*, C_1 and C_2. These measures are formulated in the case of two concepts, but they can be easily extended to more concepts.

$$C_1 = \frac{1}{2}\left(\frac{R_{l_1}\left(l_1 * l_2\right)}{R_{l_1 * l_2}\left(l_1 * l_2\right)} + \frac{R_{l_2}\left(l_1 * l_2\right)}{R_{l_1 * l_2}\left(l_1 * l_2\right)} \right) \tag{1}$$

$$C_2 = \frac{1}{2}\left(\frac{R_{l_1}\left(l_1 * l_2\right)}{R_{l_1}\left(l_1\right)} + \frac{R_{l_2}\left(l_1 * l_2\right)}{R_{l_2}\left(l_2\right)} \right) \tag{2}$$

Where $l_1 * l_2$ is the type of combination chosen (i.e., the logical operator represented by the symbol "$*$" depends on the scenario we want to test, AND $\wedge$ *or OR* $\vee$), and $R_x(y)$ is the number of retrieved images that are associated with the concept y, while relevance feedback is provided according to concept *x*; i.e., in Equation (1), $R_{l_1}\left(l_1 * l_2\right)$ is the number of images that exhibit the concept $l_1 * l_2$ among those that are retrieved by relevance feedback when the images that exhibit concept l_1 are marked as relevant.
The concept correlation measure C1, Equation (1), answers the questions proposed above: i.e., for a given combination of concepts $l_1 * l_2$, it measures if it is better to have the feedback driven by the combination rather than by the single concept. Thus, if a large value of C_1 is attained, thus it means that we can retrieve a larger number of relevant images according to the combined concept by marking as relevant just the individual concepts (separately), than marking as relevant the combination of them. If the value of C_1 is small, thus it means that the opposite holds. The concept correlation measure C_2, Equation (2), shows how much a single concept is connected with the combined concept in the relevance feedback retrieval process. For example, it can be used to see how much the combination of concepts is "embedded" in a single concept. In the case of the multiple concept AND scenario when the value of C_2 is equal to 1 it means that the retrieved images exhibits both tags l_1 and l_2, thus the two tags are highly correlated.

The "new" evaluation procedure

According to our reasoning above, we propose the following new evaluation procedure for the assessment of performances of relevance feedback techniques:

1. Test each Relevance Feedback methodology for all of the described new scenarios
2. Compare the performance attained by RF with those attained by content based browsing
3. Validate the performance by using the concept correlation measures C_1 and C_2.

It is worth noting that it is usually possible to test the first three scenarios in a multi-tagged dataset. Instead, in the case of the fourth scenario, i.e., the multiple concept AND-OR scenario, at least a combination of three tags is required, and it may happen that this condition is fulfilled only by a small set of images (i.e., is not possible to perform reliable experimentations).

With the term “content based browsing ” we mean nothing more than showing the user the n images most similar to the query with no feedback, where n is the total number of images showed to the user during the RF interactions. The aim of comparing relevance feedback with browsing is to show the benefits of relevance feedback. To put it simple: can a relevance feedback approach retrieve more relevant images than simply browsing the collection by sorting the images according to the visual similarity with the query?
Moreover, by using the above correlation measures, it is possible to perform a deeper analysis of the tested combinations by only looking at those that are valuable according to the measures proposed.

RELEVANCE FEEDBACK TECHNIQUES FOR THIS EXPERIMENTAL SET-UP

In this section, we describe the two relevance feedback techniques that we tested against the proposed scenarios . One technique is based on the nearest-neighbor paradigm, while the other technique is based on Support Vector Machines. The use of the nearest-neighbor paradigm for relevance feedback is motivated by its use in a number of different pattern recognition fields, where it is difficult to produce a high-level generalization of a class of objects (Aha et al., 1991; Duda et al., 2001). Support Vector Machines (SVM), on the other hand, are one of the most popular learning algorithms used when dealing with high dimensional spaces as in CBIR (Cristianini & Shawe-Taylor, 2000; Tong & Chang, 2001). In particular, SVM are trained by formulating the RF problem as a two-class problem. Even if the number of training samples is quite limited being related to the number of images marked by the user, SVM proved to be effective.

k-NN relevance feedback

The nearest neighbor approach to relevance feedback was proposed in (Giacinto, 2007). A score is assigned to each image of a database according to the computation of the ratio of two distances: the distance of the image from the nearest relevant image, and the distance of the image from the nearest non-relevant image. This score is further combined to a score related to the distance of the image from the region of known relevant images. In fact, in cases when very few relevant images are available, the ratio may produce a high score for all the images that are not similar to non-relevant images. Thus, we have no guarantee that images with a high score computed according to the ratio of distance, are actually relevant to the user. For this reason, when the number of relevant images is small compared to the number of images marked by the user, the score obtained by the ratio is “moderated” by a score related to the locality of the images with respect to the available relevant images. The combined score is computed as follows:

$$rel^{f}(I) = \left(\frac{n/k}{1+n/k}\right) \cdot rel_{BQS}(I) + \left(\frac{1}{1+n/k}\right) \cdot rel_{NN}(I) \tag{3}$$

where n and k are the number of non-relevant images and the whole number of images retrieved after the latter iteration, respectively. The two terms rel_{NN} and rel_{BQS} are computed as follows:

$$rel_{NN}(I) = \frac{\left\| I - NN^{nr}(I) \right\|}{\left\| I - NN^{r}(I) \right\| + \left\| I - NN^{nr}(I) \right\|} \tag{4}$$

$$rel_{BQS}(I) = \left(1 - e^{1 - \frac{d_{BQS}(I)}{\max_I d_{BQS}(I)}} \right) \Big/ (1 - e) \tag{5}$$

where *NN(I)* denotes the nearest neighbor of I, $\|\cdot\|$ is the metric defined in the feature space at hand, and d_{BQS} is the distance of image I from a reference vector computed according to the Bayes Decision Theory. This reference vector is computed so that it lies in a region of the feature space where it is more likely to find other relevant images (Giacinto & Roli, 2004). If F feature spaces are used, a score *rel(I)* is computed for each f feature space. The following combination is performed to obtain a "single" score:

$$rel(I) = \sum_{f=1}^{F} w_f \cdot rel^f(I) \tag{6}$$

where w_f is the weight associated with the f-th space. The weights w_f are estimated by taking into account the minimum distance between all the pairs of relevant images (Piras & Giacinto, 2009), and the minimum distance between all the pairs of relevant and non-relevant images as follows

$$w_f = \frac{\sum_{i \in R} d^f_{\min}(I_i, R)}{\sum_{i \in R} d^f_{\min}(I_i, R) + \sum_{i \in R} d^f_{\min}(I_i, N)} \tag{7}$$

SVM relevance feedback

Support Vector Machines can be used to estimate a decision boundary between relevant and non-relevant images in each of the feature spaces f in F.
The use of an SVM for this task is very effective because, in the case of image retrieval, we deal with high dimensional feature spaces. For each feature space f, a SVM is trained using the feedback given by the user. The results of the SVMs in terms of distances from the hyperplane of separation are then combined into to a relevance score through the Mean rule as follows

$$rel_{SVM}(I) = \frac{1}{F} \sum_{f=1}^{F} rel^f_{SVM}(I) \tag{8}$$

EXPERIMENTAL RESULTS

The reported experiments are related to three out of the four scenarios that we propose for evaluating a relevance feedback methodology in real world datasets, namely: the *single concept scenario*, the *multiple concepts AND scenario*, and the *multiple concepts OR scenario*. The aim of these experiments is to take into account a multi-tagged image dataset, and to verify the different performances of relevance feedback techniques in the proposed scenarios. The fourth scenario proposed in this paper, i.e., the *multiple concepts AND-OR scenario*, has not been considered because the available datasets did not contain a significant number of images whose combination of tags was representative of the scenario to be tested.

Dataset and experiments setup

For the purpose of testing the "new" scenarios proposed, the MIRFLICKR-25000 collection (Huiskes & Lew, 2008) has been chosen. This collection consists of 25000 images downloaded from the social photography site Flickr through its public API. The average number of tags per image is 8.94. In the collection there are 1386 tags that occur in at least 20 images. Moreover, some manual annotations are also available (24 in the collection used). We chose to represent the images in the following feature spaces:

- *Scalable Color* (Chang et al., 2001), a color histogram extracted from the HSV color space;
- *Color Layout* (Chang et al., 2001), that characterizes the spatial distribution of colors;
- *RGB-Histogram* and *HSV-Histogram* (Lux & Chatzichristofis, 2008), based on RGB and HSV components of the image respectively;
- *Fuzzy Color* (Lux & Chatzichristofis, 2008), that considers the color similarity between the pixel of the image;
- *JPEG Histogram* (Lux & Chatzichristofis, 2008), a JPEG coefficient histogram;
- *Appearance-Based Image Features* (Deselaers et al., 2008) obtained rescaling the images to 32x32 size and returning a color histogram extracted from the RGB color space;
- *Edge Histogram* (Chang et al., 2001), that captures the spatial distribution of edges;
- *Tamura* (Tamura et al., 1978), that captures different characteristic of the images like coarseness, contrast, directionality, regularity, roughness;
- *Gabor* (Deselaers et al., 2008) that allows the edge detection;
- *CEDD* (Chatzichristofis & Boutalis, 2008) Color and Edge Directivity Descriptor;
- *FCTH* (Chatzichristofis & Boutalis, 2008) Fuzzy Color and Texture Histogram.

We analyzed all the tags of the collection by a semantic point of view, and fused tags with the annotations in a tag verification process. After this process, we decided to keep only the tags that occur in at least 100 images. This decision was taken because we aimed to have the single concept adequately represented. This process of fusing and discarding tags brought us to keep 24718 images and 69 tags, with an average number of tags per image of 4.19. Then, we evaluated all the possible combination of concepts that derive from the modified collection for testing the multiple concept scenarios. The result of the evaluation was that only the pairs of concepts were worth to be used in a experimental phase, i.e., the number of terns, and higher combinations were shared by small subsets of images to be worth of being used in the experiments. Thus, for the pairs of concepts we kept 658 of them which occur in at least 25 images, for the purpose of testing the *multiple concepts AND scenario* and *multiple concepts OR scenario*. Finally, as query images we chose 1294 of them from the refined collection. These query images have a number of tags per image that varies from 3 to 10, with an average number of tags per image equal to 4.69 (thus very similar to the value in the all collection).

For each one of the 1294 query image, a relevance feedback experiment had been performed by using all the tags and pairs associated with the images as a target, i.e., given a query image, we considered, one at a time, each single tag and each pair of tags as target of the retrieval process to be refined through the relevance feedback. In this way we performed over 17201 relevance feedback experiments per relevance feedback technique. Each experimentation consists of 10 iterations: the first one is based on a nearest neighbor on all the feature spaces, and the other 9 iterations are performed using one of the relevance feedback techniques described above. At each iteration n = 20 images are "shown to the user" for marking the feedback. In the second section we already pointed out the performances of relevance feedback in the proposed scenarios could be evaluated using the standard measures. Consequently, experimental results are presented in terms of *Precision*, a modified definition of Recall, that we named "user perceived" *Recall*, and the concept correlation measures C_1 and C_2 proposed in this paper. The "user perceived" Recall is a recall measure that takes into account just the maximum number of relevant images that can be shown to the user according to the number of iterations, and the number of images displayed per iteration, and it is computed as follows

$$r_p = \frac{A \cap R}{R^*} \quad , \quad R^* = \begin{cases} R, & if < n \cdot i \\ n \cdot i, & otherwise \end{cases} \tag{9}$$

where A is the number of images at the iteration i, R is the number of relevant images in the dataset (for a given target), and n is the number of images shown per iteration.

Performances

In the case of *single concept scenario* we started from each one of the 1294 query images and tested all the possible concepts as single target. Every single concept belonging to a query image was used to simulate a user that is looking for images similar to the query according to that concept. Thus, each query image has been used as a starting example for different retrieval tasks. In this way, 6070 retrieval tasks were performed for each relevance feedback technique. We compared the retrieval results attained by employing the relevance feedback techniques described in the previous section, i.e., the k-NN based (NN in the figures), and the SVM, with the retrieval results attained by browsing. As already mentioned, the aim of comparing relevance feedback with browsing is to show the benefits of relevance feedback.
In Figure 2 we show the absolute gain of the relevance feedback techniques in terms of the average results in terms of Precision, and "user perceived" Recall with respect to browsing. Reported results show that, as the number of iterations increases, the performance attained by relevance feedback increases. If we compare the two relevance feedback mechanisms considered in this paper, it turns out that SVM exhibits the largest increasing performance power.

Figure 2: Precision and "user perceived" Recall in the case of single concept scenario.

The other scenarios tested are the *multiple concept AND scenario* and the *multiple concept OR scenario*. We remind the reader that the AND scenario is similar to the "single concept" scenario, where the single concept the user is looking for is actually a combination of concepts. It turns out that this scenario is more difficult than the previous one. For these scenarios we started from each one of the 1294 queries and tested all the 658 pairs tags as target. Thus, 11171 retrieval tasks were performed for each relevance feedback technique. Performances have been compared with the same methodology used in the previous scenario. We also added a comparison using the concept correlation measures C_1, and C_2 for the AND scenario. We recall that these measures are useful to understand if, when a multiple concept retrieval task is performed, the process is driven by the combination of concepts rather than the single concept used in the combination itself.
Figure 3 shows the absolute gain attained by relevance feedback techniques in terms of the average Precision, and "user perceived" Recall with respect to browsing. The performance measures exhibit the same behavior shown in the case of single tag. With respect to the previous scenario, the performances are lower because the targets are more difficult to "learn", but it is clear that, as the number of iterations increases, relevance feedback shows its capability in providing new relevant results.

Figure 3: Precision and "user perceived" Recall in the case of multiple concept AND scenario.

In Figure 4, the average values of the concept correlation measures C_1 and C_2 are reported for all the possible tasks with pairs of tags. All the relevance feedback techniques exhibit the same behavior. The C_1 measure, computed over all the queries and pairs of tags, achieves high value for a small part of pairs only, thus meaning that in the majority of cases the feedback have to be submitted according to the combination of the tags/concepts rather than using the single tag/concept to find images that are relevant to the combination of tags/concepts. The analysis of the values of

the C_2 measure tells that in the majority of cases the tags who compose the pair are loosely correlated in the collection considered in the reported experiments. The evaluation of the results in terms of the concept correlation measures allows us to point out that the results shown in Table 3 are reliable because we have a medium-low correlation in the majority of cases. Thus it means that the combined concept can't be easily retrieved by using just the single concept, but it is necessary to use the combined concept as feedback in the retrieval process. The values of these two concept correlation measures allow us to point out that multiple concepts are actually retrieved by feedbacks related to multiple concepts rather than by feedbacks related to individual concepts, and that relevance feedback techniques are effective also in the case of the complex tasks implemented in the AND scenario.

Figure 4: Concept correlation measures C1 and C2 in the case of multiple concept AND scenario.

Figure 5 shows the absolute gain attained by relevance feedback techniques in terms of the average Precision, and "user perceived" Recall with respect to browsing. By comparing these results with those related to the previous cases, it can be seen that the performances are higher because the targets are much easier.

Figure 5: Precision and "user perceived" Recall in the case of multiple concept OR scenario.

If we recall the discussion that we made in the second section of this paper, it is easy to see that at each interaction the images that are relevant to the user's query are those which belong either to one of the concepts or to a combination of them. Consequently, at each interaction a large number of images are marked as being relevant.
Figure 6 shows the values of the concept correlation measures C_1 and C_2 for all the possible tasks with pairs of tags in the OR scenario. All the relevance feedback techniques exhibit a similar behavior. In this case, the performances evaluated according to the C1 measure are better than those attained in the case of the AND scenario, thus meaning that in the majority of cases the feedback have to be submitted according to the combination of the tags/concepts rather than using the single tag/concept for finding images that are relevant with the combination of tags/concepts. On the contrary, in this case the values attained by the C_2 measure are very high. It is worth noting that in the case of the AND scenario, the values of the C_2 measure are bounded in the range [0, 1], while in the case of the OR scenario it is in the range [0, ∞]. Summing up, the values achieved by the C_1 measure allow drawing the same conclusions outlined in the case of the AND scenarios. On the other hand, the values achieved by the C_2 measures, for these techniques and this dataset, show that in different cases the search for images exhibiting the OR-ed concepts is mainly unrelated to the search for images exhibiting the single concept.

Figure 6: Concept correlation measures C1 and C2 in the case of multiple concept OR scenario.

CONCLUSIONS

In this paper we have proposed a new way for evaluating relevance feedback techniques against multi-tagged image datasets that are widely used in image retrieval and classification tasks. In particular, we proposed different multi-concept scenarios, compared the retrieval capability of relevance feedback against retrieval results attained by browsing the collection,, and investigated the behaviour of relevance feedback mechanisms by means of two novel concept correlation measures. Our work has been motivated by two limitations in the evaluation of performances in the Content Based Image Retrieval literature. Experiments are usually carried out either by employing an automatic procedure involving a tagged dataset, or by involving a set of real users. In the latter

case, results are often hardly significant because it is difficult to involve a large user base that allows assessing the reliability of the proposed mechanisms due to the inevitably subjective assessment of each user. On the other hand, automatic testing procedures allow performing a more rapid and extensive evaluation because the behaviour of a large number of users can be simulated, but often the attained performed cannot be deemed as being realistic because hardly a real user of a Content-Based Image Retrieval System associates explicit tags to images.
Our paper aims at bringing a contribution towards enabling a more detailed assessment of the performances of relevance feedback techniques by proposing a more "realistic" approach exploiting the availability of multi-tags image datasets. Reported experiments showed the performances of two relevance feedback mechanisms based on two classification paradigms. The proposed scenarios, result comparisons, and performance measures are to be considered as a starting point for a different way of performing and analyze the behavior of relevance feedback techniques. It is clear that this contribution is limited because it is still difficult to "catch" all the details that a human eye captures in an image. Furthermore, it is still an approximation the modelling of the user behaviour as a search for images exhibiting one of the concepts expressed by the tags.

REFERENCES

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning , 6*, 37–66.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition , 37* , 1757–1771.

Chang, S.F., Sikora, T., Puri, A. (2001) Overview of the mpeg-7 standard. *IEEE Transaction on Circuits and Systems for Video Technology, 11(6)*, 688-695.

Chatzichristofis, S.A., & Boutalis, Y.S. (2008) Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In: A. Gasteratos, M. Vincze, J.K. Tsotsos (eds.) *Lecture Notes in Computer Science: vol 5008. ICVS*, (pp. 312–322). Springer.

Chatzichristofis, S.A., & Boutalis, Y.S. (2008): Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In: *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, (pp. 191–196). IEEE Computer Society.

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys , 40* , 1–60.

Deselaers, T., Keysers, D., & Ney, H. (2008): Features for image retrieval: an experimental comparison. *Information Retrieval 11(2)*, 77–107.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* . New York: John Wiley and Sons, Inc.

Giacinto, G. (2007). A nearest-neighbor approach to relevance feedback in content based image retrieval. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval* (pp. 456–463). New York, NY, USA: ACM.

Giacinto, G., & Roli, F. (2004). Bayesian relevance feedback for content-based image retrieval. *Pattern Recognition, 37* , 1499–1508.

Huang, T., Dagli, C., Rajaram, S., Chang, E., Mandel, M., Poliner, G., & Ellis, D. (2008). Active learning for interactive multimedia retrieval. *Proceedings of the IEEE , 96* , 648 –667.

Huiskes, M. J., & Lew, M. S. (2008). The MIR flickr retrieval evaluation. In M. S. Lew, A. D. Bimbo, & E. M. Bakker (Eds.), *Multimedia Information Retrieval* (pp. 39–43). ACM.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Comput. Commun. Appl., 2* , 1–19.

Lux, M., & Chatzichristofis, S.A. (2008): Lire: lucene image retrieval: an extensible java cbir library. In: *MM '08: Proceeding of the 16th ACM international conference on Multimedia* (pp. 1085–1088). ACM, New York, NY, USA.

Piras, L., & Giacinto, G. (2009). Neighborhood-based feature weighting for relevance feedback in content-based retrieval. In *WIAMIS* (pp. 238–241). IEEE Computer Society.

Rui, Y., & Huang, T. S. (2001). Relevance feedback techniques in image retrieval. In Lew M.S. (ed.). *Principles of Visual Information Retrieval* (pp. 219–258). Springer-Verlag, London.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22* , 1349–1380.

Tamura, H., Mori, S., & Yamawaki, T. (1978): Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics, 8(6)*, 460–473.

Tong, S., & Chang, E. (2001). Support vector machine active learning for image retrieval. In *Proc. of the 9th ACM Intl Conf. on Multimedia* (pp. 107–118).

Tsoumakas, G., Katakis, I., & Vlahavas, I. P. (2010). Mining multi-label data. In O. Maimon, & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 667–685). Springer.

Zhou, X. S., & Huang, T. S. (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems, 8* , 536–544.