# Enriching image feature description supporting effective content-based retrieval and annotation

Franca Debole
Consiglio Nazionale delle Ricerche
Istituto di Scienza e
Tecnologie dell'Informazione
ISTI-CNR - NMIS Lab
Via Moruzzi, 1
Pisa, Italy
Email: franca.debole@isti.cnr.it

Claudio Gennaro
Consiglio Nazionale delle Ricerche
Istituto di Scienza e
Tecnologie dell'Informazione
ISTI-CNR - NMIS Lab
Via Moruzzi, 1
Pisa, Italy
Email: claudio.gennaro@isti.cnr.it

and Pasquale Savino
Consiglio Nazionale delle Ricerche
Istituto di Scienza e
Tecnologie dell'Informazione
ISTI-CNR - NMIS Lab
Via Moruzzi, 1
Pisa, Italy
Email: pasquale.savino@isti.cnr.it

***Abstract*—The paper describes a technique that supports efficient and effective Content-Based Image Retrieval (CBIR) in very large image archives as well as automatic image tagging. The proposed technique uses a unified representation for image visual features and for image textual descriptions. Images are clustered according to their image visual features while textual content is used to associate relevant tags to images belonging to the same cluster. The system supports image retrieval based on image query similarity, on textual queries, and on mixed mode queries composed of an image and a textual part and automatic image tagging.**

## I. Introduction

The number of digital images has increased with the wide diffusion of devices for image acquisition and transmission (such as digital cameras, smartphones, etc.). These images are stored in personal archives as well as on the web and their content can be properly exploited only if effective methods for content-based image retrieval (CBIR) become available. Typically, in the content-based Image retrieval (CBIR) approach the search is not performed at the level of the actual digital content, but rather using characteristic features extracted from its content, such as shape descriptors or color histograms in case of images. In CBIR, an exact match has little meaning, and similarity concepts are typically much more suitable for searching. The problem of similarity search arises in many other unrelated areas such as statistics, computational geometry, artificial intelligence, computational biology, pattern recognition, data mining, etc. Traditional image search methods are based on visual features [21] only or on textual descriptors associated to each image. The first approach is quite poor in terms of retrieval effectiveness, due to the semantic gap between image visual features and human concepts. Thus, in many cases users prefer textual queries, which retrieve images by using the text associated to them either directly by image creators or by using techniques which try to automatically infer a textual description (e.g. by using text which is in the same document or html page) [3], [23]. This approach may conduct to misleading results - e.g. when the text is not directly related to image content - or to make impossible the access to images which either have a poor textual description or do not have any textual description at all.

There has been substantial research in organizing large image databases, which has addressed the problem of improving retrieval efficiency. This is a problem that is of paramount importance, given the continuous growth of images available on-line. However, it is even more important to improve the quality of retrieval, i.e. the system capability of selecting the highest number of items which are relevant to the user, and to limit the number of items which are not needed. In this paper we address both issues of retrieval efficiency and retrieval effectiveness, by using image visual features in combination with textual description and by supporting the creation of clusters of similar images.

Often, images are associated with text, such as captions or tags or a textual description can be extracted from documents associated to the image. We investigate strategies to use this textual information to improve searching of image documents, using sparse text. Our aim is to use both visual features and textual descriptions and by using each one of these sources of information to guide the selection of values of the other. The proposed approach is based on the use of a single, unified Image Representation, which is composed of the representation corresponding to the image visual features and the representation corresponding to the text associated to each image. This is achieved by storing both representations in the full-text Lucene (http://Lucene.Apache.org) retrieval library enhanced with content-based image retrieval facilities [12]. We create image clusters based on image visual features only and then associate to each cluster a set of terms derived from the textual descriptions associated to images belonging to the cluster. Text Information Retrieval techniques are used to select the most relevant terms to be used as tags. Image clustering is based on image visual features only since there are several images which contain poor textual descriptions or do not contain any. By using image cluster descriptions it is possible to execute image visual searches, image textual searches, and mixed mode searches which use both visual features and text.

Textual descriptions associated to each cluster can also be used to perform automatic image tagging. Given an image, we may select the cluster(s) (possibly more than one) which is (or are) more similar to the image, and use the cluster(s) textual descriptions to determine significant tags to be associated to the image.

In this paper, we exploit the possibility of transforming a set of visual features (from MPEG-7 [2], [19], [20] in particular) extracted from an image in text form. This transformation technique, which has been presented in [12], will be briefly described later in this article. However, for now what we want to emphasize is the fact that through this approach it is possible to use a standard clustering method, such as k-mean, because the texts representing the visual features associated with images are also easy to put convertible into vectors (through the well-known vector-space model). The alternative would have been to use a clustering technique that does not require the use of vectors (such as k-medoids), which is, however, more complex and costly from the point of view of the computation.

The paper is organized as follows. Next section contains a brief state of the art. Section III provides a description of the proposed approach while Section IV illustrates the experimental prototype developed. Section V contains the preliminary results of the evaluation of the proposed technique, and Section VI concludes the work and describes further planned work.

## II. Relevant Related Work

The fundamental problem of many approaches to image retrieval is that in many cases they fail to address the user's information needs: images retrieved are "similar" to the query from the system point of view, but are not satisfactory for the user that formulated the query. There is a gap (often called the semantic gap) between image representation based on low level descriptions (either image visual features or text descriptions automatically associated to images) and the real semantic meaning of the images. According to Liu [17] there are five different approaches attempted to reduce the semantic gap. One of them consists in making use of both visual content of images and of the textual information associated to them and it is one of the issues addressed in this paper.

Since the late 80ies several methods and systems supporting content-based image retrieval have been studied and developed [10], [21]. Several commercial systems and experimental prototypes have been developed, such as QBIC (IBM Query by Image Content) [6], VisualSEEk [22], Virage's VIR Image Engine [14], and VIPER [24]. These systems make use of image visual features to determine images which are the most similar to a given user query. The most common features used are global image features such are color, texture, shapes, etc. or local features. In particular, some of the most used global features are those available in MPEG-7 [2], [19], [20] while one of the most widely used local feature is SIFT (Scale Invariant Feature Transform) [18].

Other commercial systems perform image retrieval by exploiting surrounding text, such as filenames and HTML text, as primary source of information. For example Google Image Search and Yahoo! Image Search (https://search.yahoo.com/) support image search through the use of textual queries. More recently, Google enhanced its image retrieval system supporting image retrieval based on the combination of text and image visual features (http://images.google.com).

However, as already underlined, this approach may provide unsatisfactory results, since the quality of text used to index images may be of poor quality and it does not allows one to capture the semantic similarities between image query and archived images. Recent research work [3] has attempted to improve the quality of text extracted by subdividing the text related to a given image into meaningful chunks and then measuring the relevance of each chunk for the image. The relevance measure is primarily based on the position of the text and of the image in a structured representation of the HTML page.

We would like to underline that many existing approaches only use a single source of information, either text or visual features. It is also worth noting that the approaches that make use of textual information are frequently limited to Web image search, where the amount of text that can be associated to an image is quite large, and the main issue is to select the proper text. The approach that is proposed in this paper uses both image visual features and text, but is can be used even if part of the image archive has poor textual descriptions, as may happen for images tagged by non professional end users.

There were many different attempts to improve the quality of image retrieval by using other sources of information, such are pure text or a combination of multiple types of data [5] or the PARAgrab prototype system [15] which is based on the use of visual features and textual metadata. In this paper, we extend these approaches by performing image clustering based on visual features and image textual descriptions.

Image clustering which uses similarity evaluation of visual features has been used to improve retrieval speedup, which is especially important for large image repositories, such as the Web [10]. Some of the proposed clustering techniques try to combine text and visual features [11], while others use clustering to improve visualization of search results [8]. However, the proposed techniques rely on complex representations of image and textual information, so that clustering results in a complex and time consuming task that cannot be easily used for very large image archives. Instead, the proposal we make in this paper, uses a unified representation of textual and visual features, allowing us to adopt standard and simple clustering algorithms, such as k-means.

## III. The proposed approach

Let us consider an *image archive* (IA) composed of $N$ images, each one containing two parts: the *image digital content* (IDC) and the *image textual content* (ITC). These images can be those archived by a user or by an organization, or can be images obtained by crawling the web. The proposed approach has a general value and it can be applied to many different image data sets. However, it can exploit all its advantages for large IAs, with many images having an associated meaningful text.

In this paper, we do not present any method to associate a textual description to each image. It can be provided by the creator of the image, or it can be automatically extracted by text that is recognized as related to the image. For example, if images are extracted from web pages, it is possible to use part of the text which is present in the same page [3]. However, it is also possible to deal with images with an empty textual content or whose textual content does not contain any significant information (for example, a user may associate a

textual description such as "img34", which is useless when images are searched).

We can distinguish three main phases in the proposed approach:

1) a *learning phase* where image clusters are individuated and textual descriptors are associated to each cluster,
2) an *indexing phase*, where all images in the image archive are associated to each cluster,
3) an *image search phase* where image visual features and image textual descriptions are used to support content-based image retrieval.

The *learning phase* uses a subset of the entire dataset: the Learning Set (LS). Images belonging to LS are analyzed in order to determine for each image an Image Descriptor which is used to index all images. Then, images are clustered using their visual descriptors. The textual information associated to images belonging to each cluster is used to provide a description of all images belonging to the cluster. Traditional IR techniques based on term frequency are used to select the most relevant terms to be used. The result of the learning phase is a set of clusters, each one described through a set of textual terms and a visual descriptor. Each cluster is represented by the centroid of the visual descriptors and by the list of terms associated to the images belonging to the cluster. A measure of the relevance of each term is provided through a *Normalized Term Frequency* measure which is given by the frequency of each term in the cluster divided by the number of images in the cluster.

During the *indexing phase*, for each image in the archive, we determine the cluster whose visual descriptor is the most similar to the image visual descriptor and we insert it there. The insertion of a new image in the cluster may result in an update of the cluster visual descriptor and an update of the cluster textual descriptor. The same approach is used if new images are added to the archive.

The learning phase and the indexing phase can be unified if the number of images in the IA is small. Indeed, during indexing, the cluster generation is the most time consuming activity and even for quite efficient cluster generation methods (e.g. k-means clustering as used in this paper), cluster generation has a polynomial complexity on $N$. Thus, if $N$ becomes too large cluster generation may require an unacceptable processing effort.

The image index description created so far can be used to retrieve images according to different modalities.

- We can search an image through a textual query. The search uses the text to access the cluster image descriptors and to select the cluster that best matches the query. All images belonging to the cluster are retrieved. Since each image in the cluster has also its own specific textual description, this text can be used to rank the images in decreasing relevance order.
- A visual query can be executed. The query is an image and we extract its visual descriptor. During the query execution, we may use all image visual descriptors, exactly as it is usually done in many Image DBs, or we may compare the query visual descriptor to the cluster visual descriptors. This results in significant performance improvements.
- The query is a combination of text and visual query. The query is a set of textual terms and an image. We extract visual features from the image(s) and create an image visual query descriptor. Two clusters are selected: the first matches with the textual part of the query, while the second matches with the visual part (it is possible that the two clusters coincide). Images in both clusters are ranked in decreasing relevance order and returned to the user.

### A. The Learning Phase

A subset of the images belonging to the *image archive* (IA) is used as a *Learning Set (LS)*. Images in the $LS$ can be randomply selected. However, we preferred to include in the $LS$ images with a textual information containing relevant terms.

Let us assume that $M$ images are included in $LS$. $M$ must be a representative number of elements of IA which allows one to generate a number of clusters which properly describe IA. At the same time, $M$ should be limited in order to reduce the cost of cluster generation. The best choice of $M$ requires to balance between these two conflicting criteria. In the current version of the paper, we will not investigate on techniques to properly select $M$, which is taken to be approximately $10\%$ of $N$ (we will use $M = 10^5$ for an image archive composed of $N = 10^6$ images).

The preliminary activity of the *Learning Phase* is a pre-processing of the $M$ images in order to create an index structure to be used for all successive clustering and search activities.

During pre-processing, the IDC part of each image is analyzed in order to extract an *image visual descriptor* (IVD) by using image visual features such as *color distribution*, *image texture*, *object's shapes*, etc. (we use MPEG-7 descriptors [2], [19], [20]). The proposed method may use any type of visual features, but the quality of the results will depend on the specific features selected. Our aim, as described in the future work, is also a detailed evaluation of the retrieval effectiveness in order to determine the best combination of visual features and textual descriptors to be used.

The IT is analyzed in order to generate an *image textual descriptor* (ITD). Words in the textual content are reduced to their stem and then analyzed in order to remove non relevant data (such as words which are too frequent, text in a non relevant language, text that is considered as non significant, etc.). The current version of the prototype system uses only these basic Information Retrieval techniques to analyze textual content. However, in the future we plan to improve the quality of the ITD by performing more sophisticated analysis of the text. For example, we may recognize the language text and translate all terms to a single language, or we may recognize synonyms in order to have broader textual descriptions.

The complete image descriptor (ID) is equal to ID $=$ (IVD, ITD).

The IVD is used to generate a *Surrogate Textual Representation* (STR) to the Image Digital Content (IDCs) that we have to search for similarity. This textual representation is made specifically to be treated with conventional inverted index of text-based search engines. The algorithm for generating STRs needs the following information: 1) a set of representative IDCs, called *Reference Objects* (taken for instance from the dataset that we wish to index) and 2) a distance function to assess the dissimilarity between any two IDCs. Employing the distance function and the reference objects, our algorithm accepts as input any ID and produces the corresponding STR as output. This textual representation is then used (on behalf of the original ID) for indexing and querying purposes: STRs can be indexed with a standard text-based search engine (such as Lucene, http://Lucene.Apache.org). The search can be performed by submitting the SRT corresponding to the query to the search engine and by collecting the results set as a ranked list of the IVDs associated with their SRTs, in the same way as in conventional full-text searches. Note that the distance function (or a similarity measure) is only used to produce a total order of a set of IVDs in order of decreasing similarity with a given query IVD; no scores or other indicators are required by the algorithm.

The assumption that is at the basis of this technique has been introduced by Chavez et al. [9] and elaborated in [4]. It consists on observing that if two objects $o_1$ and $o_2$ of a metric space are very similar (which in metric spaces means that they are close one to each other), their view of the surrounding objects (their perspective) is similar as well, i.e, sorted in the same way. Accordingly, we can use a measure of similarity between the orders of the surrounding IVDs, in place of the original distance function for matching IVDs. We refer to this basic idea as the *perspective based space transformation*. This measure of similarity between the ranks can be obtained from the function of similarity used by the search engines (cosine similarity) using the STRs generated from the IVDs. In this way, we are able to set up a robust information retrieval system based on well-tuned code bundled in available text-based search engines that combines full-text search with content-based image retrieval capabilities.

The $M$ images belonging to the Learning Set are clustered into *nc* clusters by using the IVD of the images. In particular the STR representation of the IVD is used. We used a k-means clustering algorithm [16] and we selected the value *nc* that maximizes the *cluster compactness* and *cluster separation* [13]. We underline that the specific image similarity measure adopted, based on SRTs similarity enables us to perform k-means clustering of the image descriptors, since images are represented as vectors whose similarity is measured through Euclidean distances. The adoption of an image representation based on the complete image descriptor (ID) would have implied that image similarity measures are more complex and k-means cannot be used.

Each cluster is represented through a *cluster descriptor* (CD) which is composed of the *cluster visual descriptor* (CVD) and the *cluster textual descriptor* (CTD). The CVD of each cluster is given by its centroid determined by through the STRs of images belonging to the cluster.

The CTD of each cluster is created by using the set of terms which are present in the images belonging to the cluster. Let us consider that i-th cluster contains $N_i$ images. Then $\text{CTD}_i = \cup_{n=1}^{N_i}\{\text{ITD}_{in}\}$ where $\text{ITD}_{in}$ is the n-th textual term of cluster *i*. In order to select the most significant terms, we measure the significance of each term as a *Normalized Term Frequency* $\text{Ntf} = \text{tf}/\text{Cf}$, where *tf* is the number of occurrences of the term in the cluster and *Cf* is the number of images in the cluster. A threshold is used to remove the less representative terms, while the others are retained together with their *Ntf*.

### B. Indexing

All remaining images of the image data set are indexed through the following procedure:

- We generate an Image Descriptor $ID = (IVD, ITD)$ for each image by using the same approach used in the pre-processing of the Learning Phase.
- The IVDs are transformed into an STR representation.
- The STR representation of each image is compared with the STR representation of the cluster centroids and the similarity between the image STR and the cluster centroid STR is computed.
- The image is inserted into the cluster having the highest similarity of STRs.
- After the insertion of all images in the image data set, the cluster centroids (both the textual and visual parts) are updated to take into account the updates.

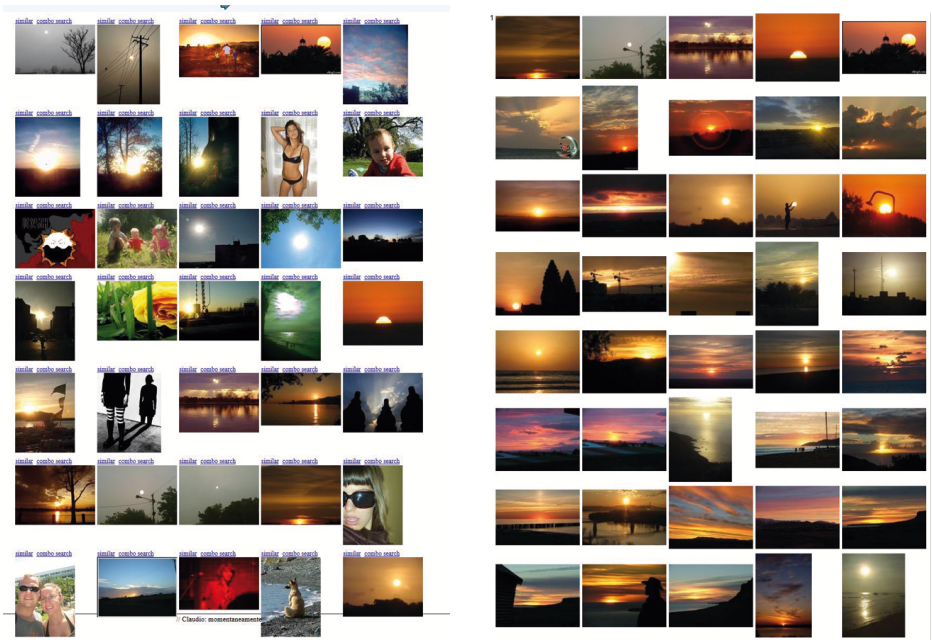
At the end of the indexing phase, the $N$ images of the Image Archive IA are grouped into *nc* clusters $C_1, ..., C_{nc}$, where a generic cluster $C_i$ contains $N_i$ images $I_{11}, ..., I_{iN_i}$, and it is represented through a cluster descriptor $\text{CD}_i = (\text{CVD}_i, \text{CTD}_i)$.

### C. Image Search

The proposed indexing method enables the execution of content based image retrieval through many different modalities. We may execute three types of queries: (i) image similarity search queries, (ii) image retrieval based on text-based queries, and (iii) mixed-mode queries which use both images and text in query formulation. The first type uses an image as query and retrieves the most similar images to the query image. The similarity between the query image and the images in the archive is measured by using the image visual features. Image retrieval through text-based queries makes use of the textual descriptors associated to images and executes a free text search query on them. The third type of query specifies a query image and a set of words. It retrieves the most similar images to a the query image and whose textual description is similar to the words specified in the query.

For each one of the three query types, When executing a query we may either access directly the entire image data set, or we may take advantage of the image clustering process performed during image indexing.

1) It is possible to execute an image similarity search on the entire data set, by using the retrieval technique proposed in [12]. Image visual features are extracted from the query image and are used to access the image STRs stored in the Lucene archive. The $k$



(a) (b)

Fig. 1. Example of the formulation of a search based on the use of a text as query (Query = ”Sun”). (a) Query executed on the entire image archive. (b) Query executed by accessing cluster centroids first.

most similar images (according to the image visual features) are retrieved.

2) We may also perform image similarity search on the entire dataset by using the clusters created during the indexing phase. Query execution requires that the image visual features are extracted from the image and that the query image STR is created. The query image STR is compared with the cluster centroids, and the cluster with highest similarity is selected. Then, we may order to images in the cluster in decreasing similarity with the query image. We expect that the results obtained are quite similar to those obtained with the previous method. However, this approach is expected to provide a faster access (since we are accessing only a single cluster instead of accessing the entire dataset).
3) Text-based queries which use the text annotations associated to images can be performed on the entire dataset by using the text retrieval functionalities offered by Lucene.
4) Text based queries may also be executed by selecting the cluster with the highest textual similarity with the query and then ordering images according to these terms. This approach is expected to provide, as for image-based queries an improvement of retrieval efficiency. It is also expected to provide an improvement of retrieval effectiveness, since allows one to retrieve images with poor textual descriptions but that belong to the same cluster of images with a textual description which matches the query.
5) A mixed mode query can be executed by first selecting the cluster that matches the image part of the query and the cluster that matches the textual part of the query. The final result set is given as a combination of the two result sets.

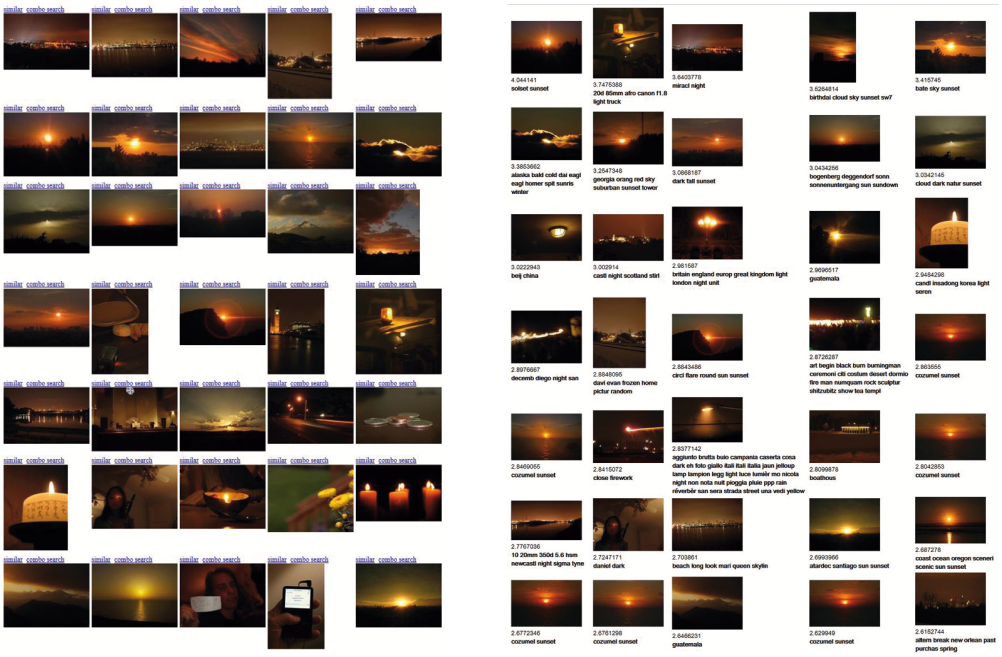
## IV. The Experimental Prototype System

We implemented an image retrieval system based on the approach described in previous section. The prototype has been implemented in Java by using Lucene as a system for text retrieval and Mahout (https://mahout.apache.org/) for clustering.

in the following we will refer to methods that compare the query with the entire data set as *Basic Method*. In Figure 1 we report the result we obtain for textual queries (Query = ”Sun”) when the *Basic Method* is used (1a), and when the query is executed by accessing cluster centroids first (1b), as proposed in this paper. We may observe that the objects in Figure 1b contain less false positive results than those present in Figure 1a.

In Figure 2 we report a comparison of the *Basic Method* and the method proposed in the paper, for an image similarity query. The first result set corresponds to a query executed on the entire image archive, while the second reports the results we obtain if cluster centroids are accessed first. We may observe that the objects retrieved are quite similar in both cases, even if the result we obtain using image clusters is slightly better.

In Figure 3 we report an example of a mixed mode query which combines the queries used in Figure 2 and Figure 1. The result set obtained exhibits a better quality if compared to those obtained with image search and text search alone. For each result we report the relevance score and a list of most relevant tags associated to the image. This is quite useful to the end user, since it allows to refine the selection. For example, it is possible to select the images with a sky (as specified in the text query and in the image query) with a sunset.



(a) (b)

Fig. 2. Example of the formulation of a search based on the use of a single image as query. (a) Query executed on the entire image archive. (b) Query executed by accessing cluster centroids first.

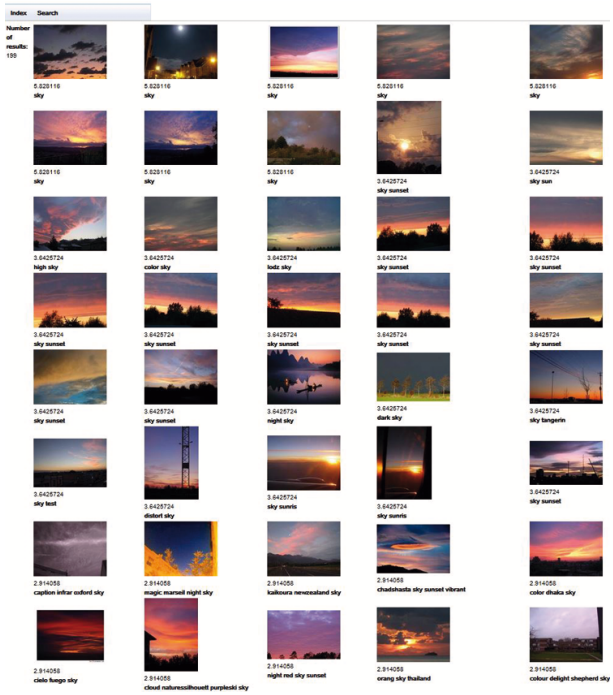


Fig. 3. Example of the formulation of a search based on the use of text and images as query.

## V. Experimental evaluation

In this section, we report the results of an experimental evaluation of the proposed method. The image dataset used for the evaluation is composed of part of the CoPhIR dataset [7] which consists of 106 millions images, taken from Flickr [1], described by MPEG-7 visual descriptors. Most of these images have an associated textual description which has been provided by the creators of the image. However, since there is no control in the tags provided by users, there are many textual descriptors of poor quality. Content based retrieval can be performed by using similarity functions of the visual descriptors associated with the images.

The learning phase has been performed on a subset of the image dataset composed of $10^5$ images selected among those with an associated textual description written in English language. In the learning phase, we performed several experiments with different numbers of clusters (*nc*). We tried with nc = 500; 1,000; 2,000; and then selected nc = 2,000 which provides the most compact clusters.

We will measure the query response time for different types of queries, and we will compare the system effectiveness with an existing image similarity retrieval system. In particular, we will compare the retrieval effectiveness of the technique proposed in this paper with the effectiveness of a system that supports retrieval based on image textual descriptions and with a system that performs image similarity search based on the use of the same image features we are using in our prototype.

The standard approach to information retrieval system evaluation is based on measuring the number of *relevant* and *non relevant* objects retrieved. Measures are conducted by using a *ground truth* collection composed of a set of documents and a set of queries which have been already evaluated in order to determine the exact number of relevant and non relevant objects. The measures used to determine the effectiveness of a retrieval system are the *precision*, i.e. the percentage of relevant objects among those retrieved, and the *recall*, i.e. the percentage of relevant objects retrieved among those that are

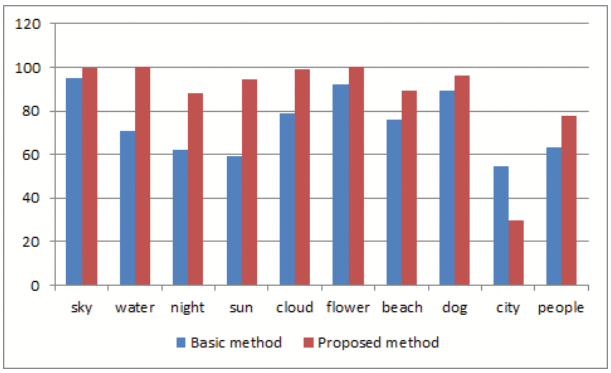



Fig. 4. Comparison between basic method and proposed method of precision at 100 for a text query. All 10 queries are shown.

relevant to the query.

However, in many applications where the number of objects in the data set is very large, it is difficult to measure both precision and recall (in particular, the measure of recall requires to know the number of relevant objects to the query). Furthermore, in many applicative settings, what really matters a user when searching in large image archives is rather how many good results there are on the first page or pages of retrieved objects. This leads to measuring precision at fixed values of retrieved results, e.g. $10, 50,$ or $100$ objects. This is called *Precision at k* (P-k), and it is the measure that we will use in this paper. It has the advantage of not requiring any estimate of the number of relevant objects, so that it can be easily measured. However, it has the disadvantage that the total number of relevant objects for a query has a strong influence on precision at k. There are other relevance measures which partially alleviate this problem. We plan to consider them for future research, but for the moment the measure of P-k is sufficient to obtain an estimate of the improvements that the proposed approach can offers.

In order to define the queries to be used for the evaluation, we defined 10 different concepts that have been used as textual queries. In order to simplify the evaluation, in the current experiments we expressed only simple queries, such as retrieve all images containing the sun, or all images with a sky. Then, we selected, for each concept, a limited number of images that represent it and that are used as visual queries. Mixed-mode queries are given as a composition of a visual query and a textual query.

The same set of queries has been used for our prototype system and for a free text retrieval system using the image textual descriptors, and an image retrieval system using the same image features. In order to limit the influence of the number of relevant results on the values of P-k, we used concepts which are quite broad so that the number of expected relevant items is quite large.

The query results are then presented to a user ($10, 50, 100$ results) which has to select those that satisfy his information needs for the specific query. The average value of the number of relevant items for all queries is then used to provide a measure of P-10, P-50, and P-100. The evaluation has been performed by 5 users, and the results were averaged. This allows us to obtain results which are not biased by the judgments of a single user. Moreover, queries which require to use an image as a query, have been repeated with different images expressing the same concept.

TABLE I. FREE TEXT QUERY

| P-k | Basic Method | Proposed Method |
|---|---|---|
| P-10 | 77.5 | 91.0 |
| P-50 | 76.1 | 86.6 |
| P-100 | 74.10 | 87.30 |

TABLE II. IMAGE SIMILARITY QUERY

| P-k | Basic Method | Proposed Method |
|---|---|---|
| P-10 | 82 | 83 |
| P-50 | 66.7 | 69.8 |
| P-100 | 62.8 | 65.4 |

TABLE III. MIXED MODE QUERY

| P-k | Basic Method | Proposed Method |
|---|---|---|
| P-10 | 84.7 | 87.5 |
| P-50 | 84.6 | 89 |
| P-100 | 80.7 | 87.9 |

Table I reports the percentage of relevant results for 10, 50, and 100 results. We may observe that the improvement of precision is of more than 10%. However, when we analyse in more detail the results for single queries, we observe that the variance of precision as measured for the 10 queries slightly increases when we use the method proposed in this paper. Thus, we may conclude that clustering results in a very significant improvement for some queries, while for other queries this is not the case.

In order to illustrate this behaviour, in Figure 4 we report a comparison of P-100 among all text queries executed by using the basic method and text queries executed by using the method proposed in this paper. We may observe that not all queries exhibit the same improvement and for one specific query, the precision is reduced with the proposed method.

Table II reports the results for the image similarity queries. The comparison is made between queries executed on the entire image data set and queries executed by using the image clusters. Again we can observe an improvement when the method proposed in this paper is used, even if the quality improvement is between 1% and 3%.

Table III reports the results for the mixed mode queries. Again the proposed method exhibits the best quality, at all P-k values, with an improvement of precision which goes to 7% for P-100.

We also observe an improvement of query response time when using the proposed method. Indeed, we are accessing only a subset of image archive during query execution.

## VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown how to use a unified representation for images based on the visual content combined

with the text associated to the image in order to improve the performance of a real CBIR system. The improvement has been reported both in terms of effectiveness that the efficiency. To achieve this, we used a technique that transforms the visual image represented by low-level features in the text, so that they can be grouped through a standard clustering algorithm such as k-means.

A number of extensions are planned as future work. We intend to improve the Learning Phase by defining appropriate criteria for the selection of the number of objects to be included in the learning set. Moreover, we will investigate the adoption of techniques to improve the quality of the textual terms to be associated to each image. Indeed, the textual descriptions are very diverse, in terms of level of detail, quality of the content, language used, etc. Thus, it is necessary to study appropriate methods that allow one to remove all non relevant text and to reduce textual descriptions to a single unified language.

We also plan to continue the work on the evaluation of the system performance, by considering the adoption of different measures for the retrieval effectiveness and by using a larger number of queries. We will also conduct experiments with larger data sets and with different data set, in order to study how the quality of results depends on the characteristics of the data set.

As a continuation of the work described in this paper, we also plan to extend the technique by supporting two new methods for the enrichment of image content description. While we currently cluster images by using their visual descriptors and then we use the associated text to determine the textual descriptions to be associated to each cluster, in the future we will consider a method where clustering is performed by using the textual descriptions first. We expect that this approach will provide an improvement of retrieval quality if detailed textual descriptions are available. Another extension will require creating two different groups of clusters, the first based on the image visual descriptions and the second based on the image textual descriptions. Relations between clusters belonging to the two groups will be determined. This approach should exploit the advantages of both approaches.

Finally, we will investigate the possible use of the technique for image annotation, i.e. automatically selecting a set of textual labels to describe the semantic content of images.

## ACKNOWLEDGMENT

This work has been supported by European funds, through the program POR Calabria FESR 2007-2013 - PIA Regione Calabria Pacchetti Integrati di Agevolazione Industria Artigianato Servizi, project ITACA (Innovative Tools for cultural heritage ArChiving and restorAtion).

## REFERENCES

[1] Flickr. http://www.flickr.com/.

[2] Mpeg-7. ISO/IEC JTC1/SC29/WG11N6828, October 2004. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[3] S. Alcic and S. Conrad. Page segmentation by web content clustering. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS 2011, Sogndal, Norway, May 25 - 27, 2011*, page 24, 2011.

[4] G. Amato, C. Gennaro, and P. Savino. Mi-file: using inverted files for scalable approximate similarity search. *Multimedia Tools Appl.*, 71(3):1333–1362, 2014.

[5] G. Amato, F. Rabitti, and P. Savino. Multimedia document search on the web. *Computer Networks*, 30(1-7):604–606, 1998.

[6] J. Ashley, M. Flickner, J. L. Hafner, D. Lee, W. Niblack, and D. Petkovic. The query by image content (QBIC) system. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, 1995.*, page 475, 1995.

[7] P. Bolettieri, A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti. CoPhIR: a test collection for content-based image retrieval. *CoRR*, abs/0905.4627v2, 2009.

[8] D. Cai, X. He, Z. Li, W. Ma, and J. Wen. Hierarchical clustering of WWW image search results using visual, textual and link information. In *Proceedings of the 12th ACM International Conference on Multimedia, New York, NY, USA, October 10-16, 2004*, pages 952–959, 2004.

[9] E. Chávez, K. Figueroa, and G. Navarro. Effective proximity retrieval by ordering permutations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(9):1647–1658, 2008.

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.

[11] B. Gao, T. Liu, T. Qin, X. Zheng, Q. Cheng, and W. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005*, pages 112–121, 2005.

[12] C. Gennaro, G. Amato, P. Bolettieri, and P. Savino. An approach to content-based image retrieval based on the lucene search engine library. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, editors, *Research and Advanced Technology for Digital Libraries*, volume 6273 of *Lecture Notes in Computer Science*, pages 55–66. Springer Berlin / Heidelberg, 2010.

[13] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering validity checking methods: Part II. *SIGMOD Record*, 31(3):19–27, 2002.

[14] A. Hampapur, A. Gupta, B. Horowitz, C. Shu, C. Fuller, J. R. Bach, M. Gorkani, and R. Jain. Virage video engine. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 188–198, 1997.

[15] D. Joshi, R. Datta, Z. Zhuang, W. P. Weiss, M. Friedenberg, J. Li, and J. Z. Wang. Paragrab: A comprehensive architecture for web image management and multimodal querying. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, pages 1163–1166, 2006.

[16] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):881–892, 2002.

[17] Y. Liu, D. Zhang, G. Lu, and W. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[19] J. M. M. Sanchez. MPEG-7: overview of MPEG-7 description tools, part 2. *IEEE MultiMedia*, 9(3):83–93, 2002.

[20] J. M. M. Sanchez, R. Koenen, and F. Pereira. MPEG-7: the generic multimedia content description standard, part 1. *IEEE MultiMedia*, 9(2):78–87, 2002.

[21] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[22] J. R. Smith and S. Chang. Visualseek: A fully automated content-based image query system. In *Proceedings of the Forth ACM International Conference on Multimedia '96, Boston, MA, USA, November 18-22, 1996.*, pages 87–98, 1996.

[23] J. R. Smith and S. Chang. Visually searching the web for content. *IEEE MultiMedia*, 4(3):12–20, 1997.

[24] D. M. Squire, W. Müller, H. Müller, and T. Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters*, 21(13-14):1193–1198, 2000.