Searching through Photographic Databases with QuickLook

Gianluigi Ciocca[a], Claudio Cusano[a], Raimondo Schettini[a], Simone Santini[b], Andrea de Polo[c], and Francesca Tavanti[c]

[a]Università degl Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
[b]Universidad Autónoma de Madrid, C/ Tomas y Valiente 11, 28049 Madrid, Spain
[c]RD multimedia laboratory, Alinari 24 ORE SpA, Largo Alinari 15, 50123 Florence, Italy

ABSTRACT

We present here the results obtained by including a new image descriptor, that we called *prosemantic* feature vector, within the framework of QuickLook^2 image retrieval system. By coupling the prosemantic features and the relevance feedback mechanism provided by QuickLook^2, the user can move in a more rapid and precise way through the feature space toward the intended goal. The prosemantic features are obtained by a two-step feature extraction process. At the first step, low level features related to image structure and color distribution are extracted from the images. At the second step, these features are used as input to a bank of classifiers, each one trained to recognize a given semantic category, to produce score vectors. We evaluated the efficacy of the prosemantic features under search tasks on a dataset provided by Fratelli Alinari Photo Archive.

Keywords: Image retrieval, image indexing, semantic gap, prosemantic features

1. INTRODUCTION

The need to retrieve visual information from large image and video collections is shared by many application domains, and a wide variety of content-based retrieval methods and systems can be found in the literature. Their capability as general purpose systems is, however, in large part limited by the a-priori definition and setting of the user's aims (e.g. target search, similarity search, category search); the set of features (visual or textual) used for image indexing; the similarity metric adopted; and the way in which the user may interact with the system in order to express his/her information needs. A survey of some of the most important techniques used in Content-Based Image Retrieval (CBIR) systems can be found in.[1]

Many of the existing systems are based on image features derived from computer vision, which can be computed directly and automatically from the images themselves. However simple content-based features could not characterize the images to the degree of generality and sophistication that is required for general-purpose retrieval. In order to come to grip with this problem and to provide satisfactory retrieval performance, different solutions were introduced in the retrieval process. One of these solutions is *relevance feedback*,[2] which relies on the interaction with the user to provide the system with examples of images relevant to the query. The system then refines its result depending on the selected images. The user's feedback provides a way to infer short term and case-specific query semantics.

Other systems explicitly extract and embed in the retrieval process semantic information about the image content through the use of automatic classification techniques.[3] These techniques can then be employed to automatically annotate the image content with keywords, which are then used for retrieval. If the underlying annotation is reliable, text-based image retrieval can be semantically more meaningful than other retrieval approaches.[4] Concept detection techniques categorize images into general categories such as city, landscape, sunset, forest, sea, etc..., using supervised classification.[5] The idea here is that meaning is provided implicitly through the classification of the training set, and it will supplement and integrate the low-level information provided by the features. One of the first attempts to integrate and compare semantic keyword and low-level features into a single CBIR framework is the SIMPLIcity system.[6] A more recent paper[7] defines a new paradigm

Gianluigi Ciocca: ciocca@disco.unimib.it, Claudio Cusano: claudio.cusano@disco.unimib.it, Raimondo Schettini: schettini@disco.unimib.it, Simone Santini: simone.santini@uam.es, Andrea de Polo: andrea@alinari.it, Francesca Tavanti: tavanti@alinari.it

Multimedia on Mobile Devices 2012; and Multimedia Content Access: Algorithms and Systems VI, edited by Reiner Creutzburg, David Akopian, Cees G. M. Snoek, Nicu Sebe, Lyndon S. Kennedy, Proc. of SPIE-IS&T Electronic Imaging, Vol. 8304, 83040V · © 2012 SPIE-IS&T · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.911976

denoted as query-by-semantic-example (QBSE) which combines a query-by-example approach with semantic retrieval

Here we present the results obtained by including within the framework of QuickLook^2 image retrieval system a new image descriptor that we called prosemantic feature vector. The framework is based on the feature vector model and support a relevance feedback mechanism. With this image descriptor we try to retain the advantages of using classifiers (the semantics obtained from the annotated training set) without the disadvantages (the restriction to the categories on which it was trained). The feature vector will be embedded into a feature space in which similarity is defined as a function of the distance, and the search is done using relevance feedback. By exploiting the relevance feedback mechanism, the user can move through the prosemantic feature space toward the target image.

The search experiments are performed on a dataset of about 3,000 images provided by Fratelli Alinari Photo Archive. The prosemantic features are evaluated on this dataset exploiting the QuickLook^2 image retrieval system.

2. RELEVANCE FEEDBACK RETRIEVAL

The QuickLook^2 system allows the user to search an image data base with the aid of sample images, or a user-made sketch, and/or textual descriptions. The system's response can then be progressively refined by indicating the relevance, or non-relevance of the items retrieved. In particular, by exploiting the statistical analysis of the image feature distributions and of the textual descriptions of the retrieved items the user has judged relevant, or not relevant, the system is able to identify what features the user has taken into account (and to what extent) in formulating his judgment. It then modifies accordingly the impact of the different visual and textual features in the overall evaluation of image similarity, as well as in the formulation of a new single query representing the user's information needs.

Let $\mathbf{x}_I$ be the representation of the image I. Images can be described by different features so $\mathbf{x}_I$ is composed of different numerical vectors, each one representing an image characteristic (e.g. color histogram, shape, prosemantic features, etc...). We indicate these vectors for image I as $\mathbf{x}_I^{(1)}, \mathbf{x}_I^{(2)}, \ldots, \mathbf{x}_I^{(p)}$. Given a query Q and a image I, the dissimilarity between the two representations is computed as a weighted Euclidean distance:

$$D(Q,I) = \frac{1}{p}\sum_{f=1}^{p} D^{(f)}(\mathbf{x}_Q^{(f)}, \mathbf{x}_I^{(f)}) w^{(f)}, \tag{1}$$

where $D^{(f)}$ and $w^{(f)}$ are the dissimilarity metric and the weight associated to the feature f respectively. The weights $w^{(f)}$ allow to tune the contribution of each features in the overall similarity measure. The weights are determined according to the images selected by the user, and the query Q is computed by the query refinement algorithm. The dissimilarities are computed between the query and each image in the database. The images most similar to the query are presented to the user sorted by decreasing similarity. For a more detailed description of the QuickLook^2 system, the reader can refer to[8] and.[9]

The basic idea of the relevance feedback mechanism is that the distribution, in the feature space, of the images that the user has judged relevant (or not relevant) can be used to determine what features the user has taken into account (and to what extent) in formulating this judgment. With this information, one can accentuate the influence of the relevant features in the overall evaluation of image similarity, as well as in the formulation of a new query. The structure of the relevance feedback mechanism is entirely description-independent, that is, the index can be modified, or extended to include other features without requiring any change in the algorithm as long as the features can be expressed as numerical vectors. The relevance feedback algorithm works as follows: let R_+ the set of relevant images and R_- the set of non relevant images. The feature weights are computed as:

$$w^{(f)} = \begin{cases} \frac{1}{\epsilon} & \text{if } \|R_+\| < 3 \\ \frac{1}{\epsilon+\mu_+^{(f)}} & \text{if } \|R_+\| \geq 3 \text{ and } \|R_-\| = 0 \\ \frac{1}{\epsilon+\mu_+^{(f)}} - \alpha\frac{1}{\epsilon+\mu_*^{(f)}} & \text{otherwise} \end{cases}, \tag{2}$$

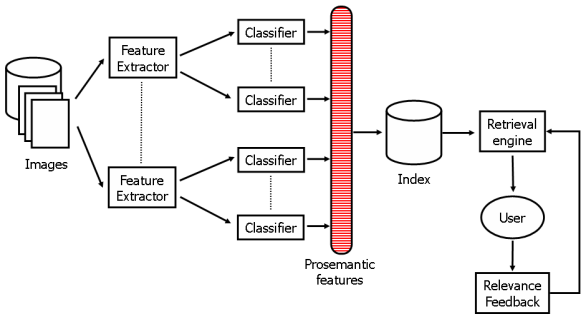

Figure 1. Prosemantic features extraction.

where ϵ and α are positive constants, $\mu_+^{(f)}$ is the average of the dissimilarities computed on the f-th feature between each pair of images in R_+, and $\mu_*^{(f)}$ the average of the dissimilarities computed on the f-th feature between each image in R_+ and each image in R_-. Negative weights are set to 0. A weight is large if the corresponding feature is present in all the relevant images while it is small or dampened if the corresponding feature assumes a broad range of values within the relevant images or if is also present in the non relevant images (viz. it is present in the relevant images but it is not relevant).

In content-based retrieval images are sometimes considered relevant because they resemble the query image in just some limited sense related to low-level features that are particularly prominent, even if semantically not very significant. Consequently, after an initial query, a given image may be selected by the user as relevant because it has one of the characteristics of the query (e.g. the same color), and another be selected for another characteristics (e.g. the shape), although the two are actually quite different from each other. To cope with this problem a method called *query refinement* is used to compute the query vector. On the basis of the images selected by the user, the system formulates a new query that better represents the images of interest to the user, taking into account the features of the relevant images, without allowing any one particular feature value to bias the query computation. Let $\mathbf{x}_I^{(f)}(k)$ be the k-th value of the f-th feature of image I. By considering only the images in the relevant set R_+, the query Q is computed as:

$$Y_k^{(f)} = \{\mathbf{x}_I^{(f)}(k) : \mid \mathbf{x}_I^{(f)}(k) - \mathbf{x}_{\bar{Q}}^{(f)}(k) \mid \leq 3\sigma_k^{(f)}\}, \tag{3}$$

$$\mathbf{x}_Q^{(f)}(k) = \frac{1}{\|Y_k^{(f)}\|} \sum_{\mathbf{x}_I^{(f)}(k) \in Y_k^{(f)}} \mathbf{x}_I^{(f)}(k), \tag{4}$$

where $\bar{Q}$ is the average query and $\sigma_k^{(f)}$ is the standard deviation of the k-th values in the f-th feature. The query is thus computed from the feature values that mostly agree with the user selection, while the outliers are removed from the computation.

3. PROSEMANTIC FEATURES

Figure 1 shows the process of the prosemantic features extraction. Prosemantic features extraction begins by describing the images with a suitable set of "low-level" features. As low level features we considered: color mean and standard deviation of the values of the LUV color channels on 9 image subregions, global color histogram in the RGB color space, statistics about the direction of edges and the descriptors associated to the Scale Invariant Feature Transform.[10] More details on how these feature have been computed can be found in.[11,12]

In order to provide a semantically meaningful information about the content of the images, each feature is used as input to an array of 14 soft classifier, trained to recognize partially overlapping classes. We selected a set of 14 classes: animals, city, close-up, desert, flowers, forest, indoor, mountain, night, people, rural, sea, street, and sunset. Some classes describe the image at a scene level (city, close-up, desert, forest, indoor, mountain, night, rural, sea, street, sunset) other describe the main subject of the picture (animals, flowers, people). The set of classes is not meant to be exhaustive, or to be able to characterize the content of the images with sufficient

specificity for our purposes. Our intent, here, was to select a variegated set of concepts providing a wide range of low-level descriptions of typical scenes. The fact that the categories are overlapping is a practical choice reinforced by an intuition based, in turn, on an analogy. As a matter of praxis, it would be problematic, next to impossible in fact, to find a reasonably extended collection of categories that show no overlap, not in the least because the very concept of "semantic overlap" is all but well defined, and can be used, at best, as a generic regulative principle.

In order to collect suitable training samples for the classifiers, we queried various image search engines on the web with several keywords related to the classes, and downloaded the resulting pictures. The images were then manually inspected in order to remove those that did not belong to the classes as well as low quality images. For each class, a set of negative examples was also selected by taking pictures from the other classes. Since the classes may overlap, a manual inspection was needed to verify that all the selected images were actually negative examples.

For each combination of low-level feature and class, a Support Vector Machine (SVM) with a Gaussian kernel has been trained. There are two parameters that need to be tuned (the cost parameter C and the scale of the Gaussian kernel γ), and they have been selected by maximizing the cross validation performance of the resulting classifier.

At the end of training, we have a distinct SVM for each feature and for each class. Given a new image Q, represented by the feature vector $\mathbf{x}_Q^{(f)}$, the SVM provides a score $s^{(c,f)}$:

$$s^{(c,f)}(\mathbf{x}_Q^{(f)}) = b^{(c,f)} + \sum_{I \in T^{(c)}} \alpha_I^{(c,f)} y_I^{(c)} \exp\left(-\gamma^{(c,f)} \|\mathbf{x}_I^{(f)} - \mathbf{x}_Q^{(f)}\|^2\right), \tag{5}$$

where $T^{(c)}$ is the training set for class c, $\mathbf{x}_I^{(f)}$ denotes the feature vectors computed on the image I, $y_I^{(c)}$ is the label in $\{-1, +1\}$ which indicates whether I is a positive or a negative example, $b^{(c,f)}$ and $\alpha_I^{(c,f)}$ are the parameters determined by the training procedure, and $\gamma^{(c,f)}$ is the scale parameter of the kernel. The score is expected to be positive when the image belongs to the class c, and negative otherwise.

Packing together the 56 scores we obtain a compact vector of prosemantic features that we place in a suitable metric space in order to index the images using relevance feedback.

4. EXPERIMENTAL RESULTS

A quantitative evaluation of the effectiveness of the prosemantic features for target search task can be found in our previous paper.[11] Here we are interested in a qualitative evaluation of the proposed prosemantic features "on-the-field" with the contribution of Fratelli Alinari Photo Archive* that supplied us a subset of 3,000 images extracted from their extensive photo archives.

Founded in Florence in 1852, Fratelli Alinari Photo Archive (now Alinari 24 ORE SpA, part of IL SOLE 24 ORE group) is the oldest firm in the world working in the field of photography, the image and communication. The birth of photography and the story of the Firm go hand in hand in their development and growth, as attested by the Alinari owned fund of 5,500,000 photographs, collected in the Alinari Archives. Alinari is a leader in Photographic Publishing, and its Art Printworks or Art Printworks is the only one in the world still using the artisan technique of collotype on paper and on silver plate from photographic images. Today Alinari is constantly updating its on-line photographic repository with over 300,000 images. An important work for long term preservation, storage and image permanence is ongoing, thanks also to an expert team of skilled technicians and photographic experts.

The dataset of images supplied to the University of Milano-Bicocca represents a new corpus of color photographs collected by the Alinari team through new photographic campaigns, in order to enrich the traditional 19th and early 20th century "vintage" repository with fresh content representing the new photographic genres and tendencies of the 21st century. In particular, the dataset is composed of 3,000 images (b/w and color), mainly regarding cityscape, landscape, art, painting and sculptures. This dataset has been taken in the center of Italy

*http://www.alinari.com/

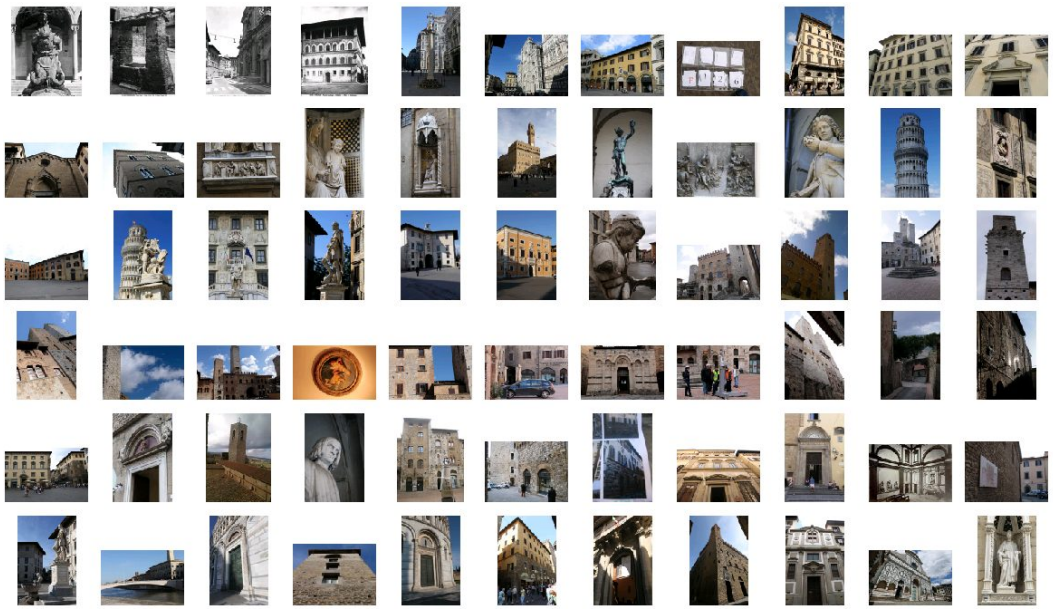

Figure 2. A small excerpt of the Alinari image dataset.

(Tuscany region) from locations of high cultural interest such as Firenze, Pisa, and San Giminiano among others. The images depict sculptures, palaces, plazas, and various other artifacts taken from different perspectives and sometimes under different illuminations. Some images represent a single object captured from a distance, while other images of the same subject have been taken much more closer (close-up). In some instances, whole panoramas of the surroundings have been acquired as well. Figure 2 shows a random selection of the images in the dataset. This dataset has been selected in order to evaluate how QuickLook^2 can perform with a specific genre, with a mix of contemporary color and historical b/w images, and to test if and how it can find images of specific objects or part of them.

The experiments have been conducted at the Alinari Archives where a copy of QuickLook^2 has been installed. Users have been asked to perform queries without supplying them with a specific task: they were free to choose the type and aim of their searches. For the purpose of the experiments, image examples are selected only from the first page of the retrieved results.

Figures 3, 4, 5, and 6 show an example of image retrieval on the Alinari dataset with the QuickLook^2 system exploiting the prosemantic features. Figure 3 shows the initial page. At the beginning the system shows the pictures sorted according to the identification code assigned by Alinari. As a result, the initial page is composed of black and white pictures only. The objective in this example was to retrieve images showing vertical sculptures, bas-relief or portraits. Thus, the query corresponds to a composite category search. It must be pointed out that the query is a very challenging one since the prosemantic features contains no information whatsoever about the concepts of "vertical", "sculptures" ad so on. Since these concepts are not directly embedded in the features, they must be indirectly deduced by the system from the image examples and the relevance feedback mechanism.

Figure 4 shows the positive/relevant images (with green bounding box) and the negative/non relevant images selected by the user. Negative images corresponds mostly to buildings while the positive images are of sculptures. It should be noted that only a subset of the available positive images and negative images in the page have been selected. Figure 5 shows the results of the first relevance feedback iteration. It can be seen that now the retrieved images are both the black and white and color ones. More images satisfying the initial query have been found, while the negative images are greatly reduced in number. The large number of different relevant images allows the user to strengthen his idea of the query by selecting images having different pictorial characteristics. This is shown in Figure 6 where the results obtained after another iteration are more coherent with the query as the relevant images are ranked in the first rows.

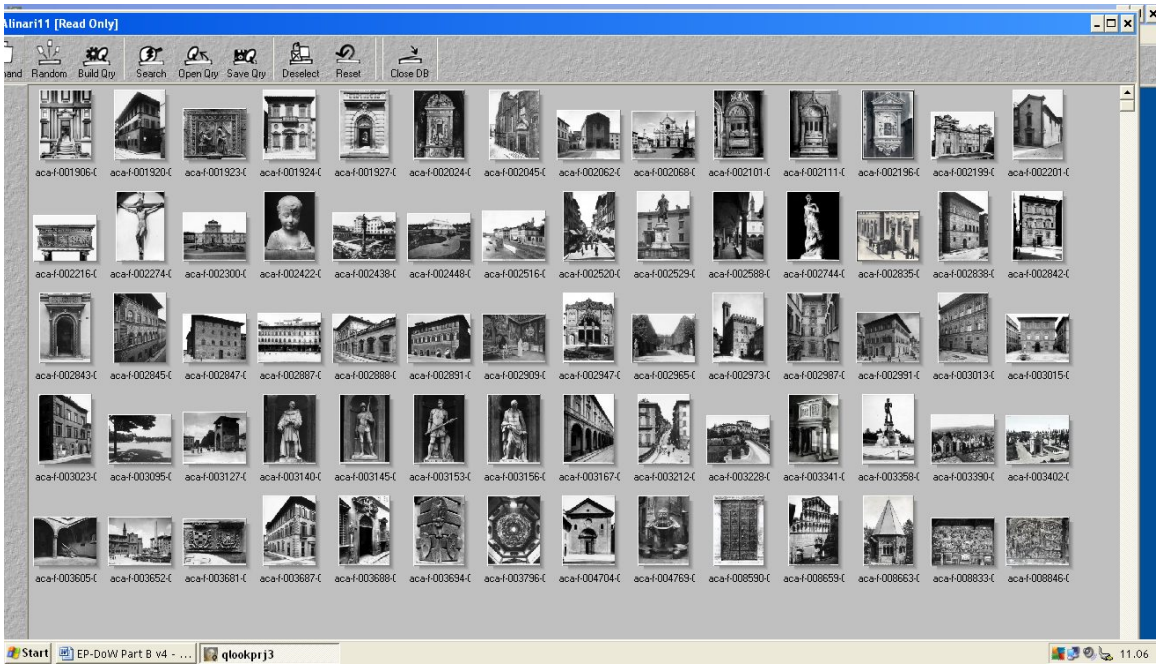

Figure 3. Initial page.

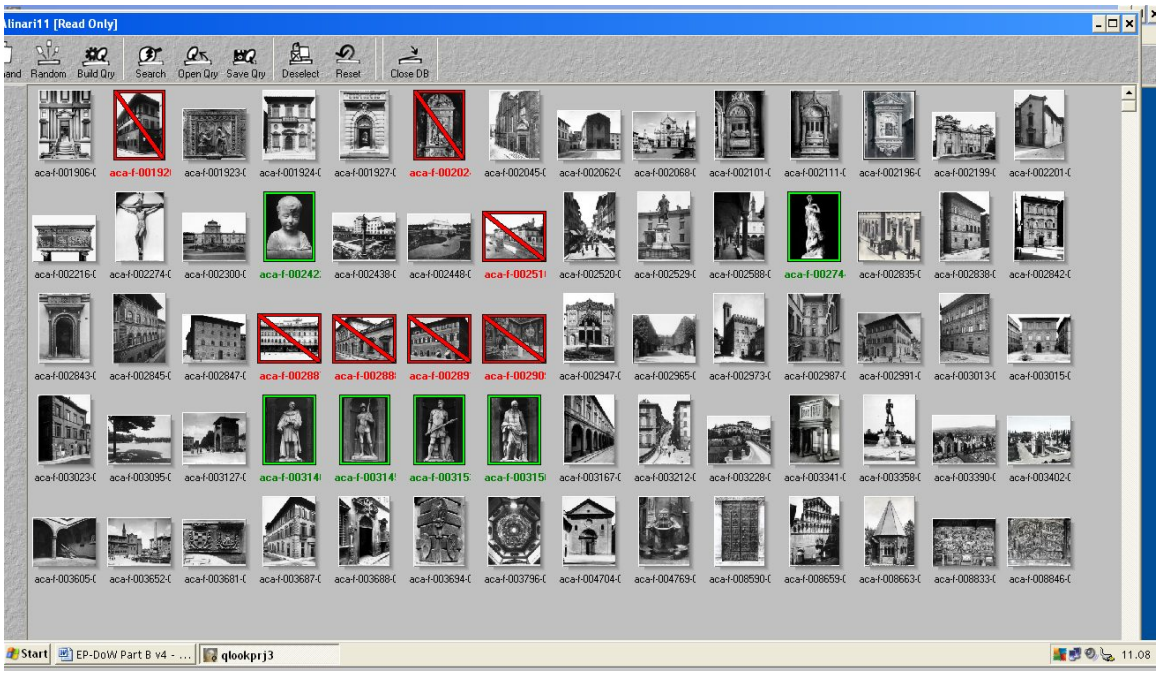

Figure 4. Selection of the positive and negative image examples for the "vertical sculptures, bas-relief or portraits" query.

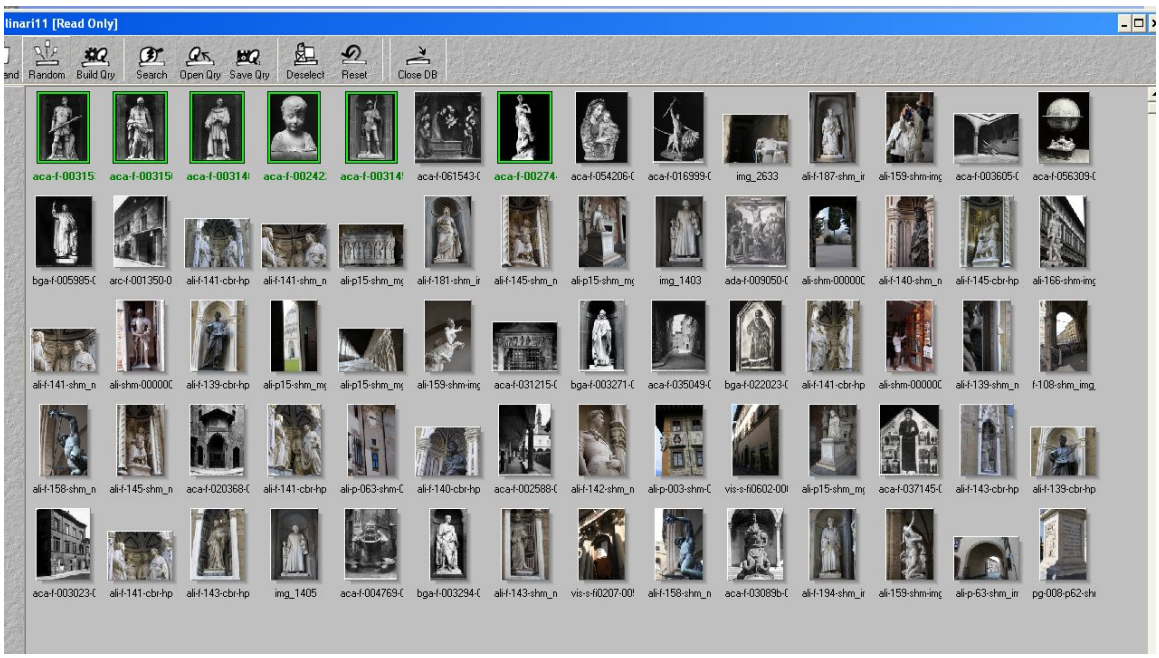

Figure 5. Results after the first relevance feedback iteration.

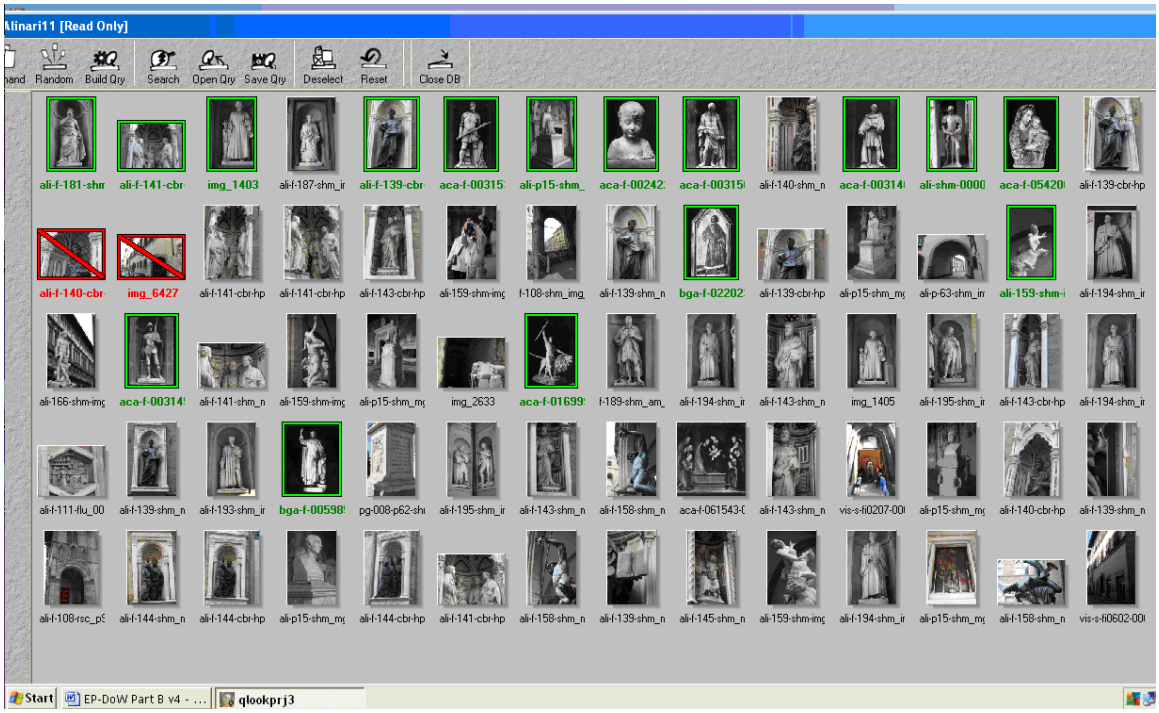

Figure 6. Results after further refinements through relevance feedback iterations.

Table 1. The usability questionnaire administered to the QuickLook^2 users. The numerical scale goes from 'strongly disagree' to 'strongly agree' (1 → 'strongly disagree', 2 → 'disagree', 3 → 'neutral', 4 → 'agree', 5 → 'strongly agree').

#	Statement	1	2	3	4	5
1.	I think that I would like to use this system frequently					⋆
2.	I found the system unnecessarily complex			⋆		
3.	I thought the system was easy to use		⋆			
4.	I would need the support of a technical person to be able to use this system					⋆
5.	I found the various functions in this system were well integrated				⋆	
6.	I thought there was too much inconsistency in this system		⋆			
7.	I would imagine that most people would learn to use the system very quickly			⋆		
8.	I found the system very cumbersome to use			⋆		
9.	I felt very confident using the system				⋆	
10.	I needed to learn a lot of things before I could get going with this system		⋆			
11.	The queries are executed rapidly					⋆
12.	Too many iterations are required to obtain acceptable results		⋆			
13.	The system allows effective category searches				⋆	
14.	The system allows effective target searches				⋆	
15.	The system is useful for browsing images					⋆
16.	The system is useful for the retrieval of images					⋆
17.	The system user interface is easy to understand	⋆				
18.	The relevance feedback mechanism is too complicated	⋆				
19.	I think that retrieval by image examples is useless	⋆				
20.	Too many positive examples must be selected		⋆			
21.	Too many negative examples must be selected		⋆			

Different users tested the system for a few days. A questionnaire in two parts was administered to these users in order to collect their impression about the system on the overall and on its functionalities. The first part was inspired by the System usability Scale (SUS) questionnaire developed by John Brooke at DEC (Digital Equipment Corporation).[13] It is composed of statements related to different aspects of the experience, and the subjects were asked to express their agreement or disagreement with a score taken from a Likert scale of 5 numerical values: 1 expressing strong disagreement with the statement, 5 expressing strong agreement and 3 expressing a neutral answer. The second part of the questionnaire focuses more on the functionalities of the QuickLook^2 system and was administered with the same modalities.

The results of the questionnaire are reported in Table 4. The score given by the users are summarized by majority vote. The results show that on the overall the system perform well. The retrieval capabilities of QuickLook^2 are judged positively as well as its efficiency and efficacy "The system is robust, fast to manage, and speedy in the query mechanism". The relevance feedback mechanism coupled with the prosemantic features is efficient and is able to retrieve satisfactory results in few iterations and requiring the users to select a moderate amount of image examples. Retrieval by examples is still considered a plus in a retrieval engine that cope mainly with images. With respect to the users' experiences the weakest point of the system is the graphical user interface. Although the selection of the images is quite intuitive, the other components of the user interface has been rated poorly. The QuickLook^2 system has been designed with a very rough and simple interface. In particular users complained about the selection of the query options and proposed some solutions such as:

- "There is not a logic approach to perform a query as the approach is through drop menu and windows and not through, for example, a wizard navigation tool. The result is that a non experienced user could have difficult at the beginning to understand where to start, which bottom to push first, and what to do, or where to go next."

- "Probably a wizard that could assist the user on the various operation, along with a graphic interface probably more colorful, without drop down menus, but with simple and direct buttons, and with an on-line help and a virtual assistant tool could push the usability."

The other recommendations can be summarized as follows: being able to save specific queries, being able to personalized the access (different users can have different privileges or access different options in the menu), have a statistic function that can send to the system developer institution a log with the bugs/system performance (for remote tracking, performance analysis, improvements). An eye tracking functionalities to detect the user behavior and evaluate what the user see, where he normally goes or select into the main windows, or even the time that a user spend between two different queries can provide vital information for future system improvements. Being able to export the query content into a virtual 3D exhibition space, or even into the Cloud, might provide additional usability possibilities.

5. CONCLUSIONS

In this paper we presented an experimentation where prosemantic features are used within the QuickLook^2 image retrieval system to perform content-based searches in a photo archive provided by Fratelli Alinari.

Employers of Alinari qualitatively evaluated the performance of the system in different retrieval scenarios. The resulting judgments are quite surprising. No criticisms have been raised concerning the system's capability in identifying the target concepts or pictures. The only negative observations were related to the user interface, an aspect of the system which was not the focus of the experimentation. In our future work we plan to thoroughly address this issue by providing a more user friendly interface, and by investigating the use of more complex visualization techniques.

REFERENCES

[1] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R., "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12), 1349–1380 (2000).

[2] Zhou, X. S. and Huang, T. S., "Relevance feedback in image retrieval: a comprehensive review," *Multimedia Systems* **8**(6), 536–544 (2003).

[3] Fan, J., Gao, Y., Luo, H., and Xu, G., "Automatic image annotation by using concept-sensitive salient objects for image content representation," in [*Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*], 361–368 (2004).

[4] Datta, R., Li, J., and Wang, J. Z., "Content-based image retrieval: approaches and trends of the new age," in [*Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*], 253–262 (2005).

[5] Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H.-J., "Image classification for content-based indexing," *IEEE Trans. on Image Processing* **10**(1), 117–130 (2001).

[6] Wang, J., Li, J., and Wiederhold, G., "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(9), 947–963 (2001).

[7] Chang, E., G., K., Sychay, G., and Gang, W., "CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Transactions on Circuits and Systems for Video Technology* **13**(1), 26–38 (2003).

[8] Ciocca, G., Gagliardi, I., and Schettini, R., "Quicklook^2: An integrated multimedia system," *Journal of Visual Languages & Computing* **12**(1), 81–103 (2001).

[9] G. Ciocca, I. G. and Schettini, R., "Content based image retrieval and video retrieval using the QuickLook search engine," *MultiMedia Information Retrieval* , 151–170 (2004).

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

[11] Ciocca, G., Cusano, C., Santini, S., and Schettini, R., "Halfway through the semantic gap: prosemantic features for image retrieval," *Information Sciences* **181**(22), 4943–4958 (2011).

[12] Ciocca, G., Cusano, C., Santini, S., and Schettini, R., "Prosemantic features for content-based image retrieval," in [*7th International Workshop on Adaptive Multimedia Retrieval*], (2009).
[13] Brooke, J., "SUS: A Quick and Dirty Usability Scale," in [*Usability Evaluation in Industry*], Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., eds., Taylor & Francis., London (1996).