Multimedia Systems manuscript No.
(will be inserted by the editor)

Miniature illustrations retrieval and innovative interaction for digital illuminated manuscripts

Daniele Borghesani · Costantino Grana · Rita Cucchiara

Received: date / Accepted: date

Abstract In this paper we propose a multimedia solution for the interactive exploration of illuminated manuscripts. We leveraged on the joint exploitation of content based image retrieval and relevance feedback to provide an effective mechanism to navigate through the manuscript and add custom knowledge in the form of tags. The similarity retrieval between miniature illustrations is based on covariance descriptors, integrating color, spatial and gradient information. The proposed relevance feedback technique, namely Query Remapping Feature Space Warping, accounts for the user's opinions by accordingly warping the data points. This is obtained by means of a remapping strategy (from the Riemannian space where covariance matrices lie, referring back to Euclidean space) useful to boost the retrieval performance. Experiments are reported to show the quality of the proposal. Moreover the complete prototype with user interaction, as already showcased at museums and exhibitions, is presented.

Keywords Image Retrieval · Relevance Feedback · User Interaction

1 Introduction

Cultural heritage preservation, valorization and exploitation has a key role in human culture. For this reason we believe this is also one of the most important areas in which multimedia retrieval research can be beneficial. All the plurality of masterpieces can be effectively enclosed into a common "paradigm" through digitalization, and augmented with multimedia retrieval techniques. This allows a significant reduction in management costs, an enormous impact on the

Daniele Borghesani · Costantino Grana · Rita Cucchiara
Università degli Studi di Modena e Reggio Emilia
via Vignolese 905b
Tel.: +39-059-2056265
Fax: +123-45-678910
E-mail: name.surname@unimore.it

public and, at the same time, a tremendous freedom of data elaboration. In other words, if wisely designed and implemented, this paradigm is able to guarantee pleasure for the regular users and usefulness for experts, providing a positive impact on society, culture and cultural heritage itself.

The ability of modern multimedia systems to account users' expectations is usually accomplished by including the user itself in the retrieval loop. The system can therefore capitalize the interaction as a primary source of information to improve the retrieval and to provide personalized content. Following the same idea, we propose a combination of content-based image retrieval (CBIR) with an effective relevance feedback strategy to create a visually assisted tagging mechanism, enclosing all these functionalities into an appealing user interface. The specific artwork for which we propose this application, that is illuminated manuscripts, is particularly interesting from the scientific point of view: many image analysis algorithms are involved, and handmade miniature illustrations provide a challenging dataset to test state of the art approaches. This context is also particularly interesting from the social and artistic point of view, since these artworks are usually inaccessible to users due to their value and delicacy. In this paper, we focus on Renaissance illuminated manuscripts, of which Italy has significant collections. Masterpieces like the *Bible of Borso d'Este* in Modena, the *Bible of Federico da Montefeltro* in Rome and the *Libro d'Ore of Lorenzo de' Medici* in Florence cannot be physically accessed, and the thousands of miniature illustrations they contain, with their story, meaning and artistic value, remain locked. The availability of a digital counterpart, augmented with all the tools of multimedia retrieval, is therefore undoubtedly desirable.

As a first contribution of this work, we highlight the benefits of covariance descriptors in the context of similarity retrieval with miniature illustrations. This feature, usually employed in literature for object detection and texture classification, provides a robust and parameter-free description capable of merging color and texture information and their correlations, and revealed particularly effective with the dataset in use. Secondly, we propose a novel relevance feedback strategy based on query reformulation and feature space warping, working nicely in the Riemannian space of covariance matrices by means of a remapping. Accordingly, we call this approach Query Remapping Feature Space Warping (QRFSW). We finally consider the joint use of visual similarity, relevance feedback and interactivity as a minor nevertheless significant novelty in the context of cultural heritage: this combination allows an innovative management of this kind of artistic content, allowing the user to browse digitally illuminated manuscripts and easily categorize miniatures using tags.

The structure of the paper proceeds as follows: Section 2 reports an analysis of the state of the art techniques in the field of interest, while Section 3 summarizes the proposed system architecture. Section 4 provides a detailed description of the covariance feature that empowers the visual search engine, while in the next Section a formalization of the relevance feedback strategy is proposed. An experimental evaluation of the methods introduced in this paper

is available in Section 5. We also included Section 6 to detail the user interface we designed for this application. Some concluding remarks and hints about future developments are finally drawn.

2 Related work

This work falls within the topic of document analysis applied to artistic material, which is a challenging research context. There is a consistent amount of image analysis and content understanding problems to solve in this field. Moreover the unusual peculiarities of this kind of visual data can lead to unpredicted and surprising results, even with state-of-the-art methods. Regarding the specific field of application of illuminated manuscripts, a process of page segmentation is usually necessary to extract from the page the relevant content, where the concept of relevant is tailored to the specific application scenario. A general but comprehensive survey of document classification has been proposed by Chen and Blostein (2007), highlighting different techniques based on image features, physical layout features, logical structure features and textual features. One of the most studied topics is obviously text segmentation and recognition. Nevertheless, in most cases a custom algorithm is required for every class of manuscripts, since the handwritten text varies a lot among styles, cultures and languages. Other proposals concentrate on specific details of the artistic work of interest, as for example drop cap letters (Coustaty et al 2011) in illuminated manuscripts. In a prior work (Grana et al 2010), we proposed a custom solution to remove text and decorations of the page. We formulated it as a classification problem, employing circular statistics to model the autocorrelation matrix in a block-wise manner. This method allowed us to highlight relevant miniature illustrations, removing background, text and decorations which were not relevant in our case. The dataset we used throughout the manuscript is the result of the aforementioned segmentation process, with an additional manual effort of correction and refinement.

At this point, the problem is reduced to a typical content-based image retrieval (CBIR), sharing the same architecture as many other in literature. In particular, the process of feature extraction is the first fundamental step in any CBIR system (Datta et al 2008), strongly impacting the quality of visual features in any other subsequent image analysis tasks, such as similarity ranking, concept detection and annotation.

A global approach, like color histograms, can provide a compact and computationally effortless summary of the content, by aggregating some information extracted at every pixel location. Instead a local approach concentrates on interest points or regions solely, to be detected and conveniently described. By designing those descriptors with scale-invariant, illumination-invariant and affine-invariant properties, a very effective tool for image matching and object detection emerges. But for more general retrieval purposes, these descriptors are usually packed into a vocabulary by means of a clustering procedure, and

the image is therefore represented again as an histogram, this time of local cues.

The reader can easily refer to an endless literature on the topic (Dance et al 2004; Mikolajczyk and Schmid 2004, 2005; van de Sande et al 2010), however it will be much more troublesome to find out the same amount of literature dealing specifically with cultural heritage and artistic content in general. Several motivations can be outlined to explain the existence of this sort of "scientific niche". Artistic content is almost always subjected to licensing and copyrights constraints, which at the best of our knowledge has prevented so far the creation of public benchmarks. Moreover it is very difficult to generate a consistent dataset to be used for testing: different artists, in different historical periods, have used very different expressive strategies, changing not only the content of their artworks but also the stylistic primitives they use to represent the content. Often the semantic gap is much more profound, given the great deal of interpretation and symbolism that artistic works usually have. There is finally a lot of emotion involved around this data, and the decision of "what is similar to what" in a set of artistic pictures could potentially go beyond the raw visual similarity up to involving very subjective statements about feelings and moods.

Recently, Hurtut (2011) proposed a comprehensive survey of content-based retrieval for artistic images. In particular, the author provided a useful taxonomy:

1. In the *image space*, we analyze the perceptual primitives (color, texture, edges), the geometric primitives (strokes, contours, shapes), the spatial arrangements of objects and regions and the semantic units (names of the objects).
2. In the *object space*, we analyze the 3D relations between object in the scene or the contextual information (illumination, shadows)
3. In the *abstract space*, we analyze the cultural aspects that can be inferred using artistic knowledge (for example the artistic style), the emotional response evoked by the artwork to the people, or the technical information about who made the artwork (drawing tools, authenticity)

In particular, our work falls into the first class of this taxonomy.

Lay and Guan (2004) proposed a painting retrieval engine based on the generative grammar of elemental concepts based on color. Corridoni et al (1998) described a retrieval strategy exploiting the sensations paintings convey, such as warmth, harmony, and contrast, according to Itten's theory. Similarly, Marchenko et al (2007) exploited brushwork and color to create an ontology, in which transductive inference was used to perform retrieval. Zirnhelt and Breckon (2007) proposed an approach based on color and texture, while Luszczkiewicz and Smolka (2009) used bilateral filtering and GMM. SIFT descriptors have been used to efficiently retrieve duplicates (Valle et al 2006), compared with Color Coherent Vectors, but limited on photographic images (not painted). In this context, the literature evidenced the lack of an ideal set of features, unlike what happens in scene recognition with natural images

where local descriptors using the bag-of-words model is known to work quite well indeed. Yen et al (2006) tried to solve the problem using Adaboost for feature selection over a set of texture and arrangement features, exploiting also relevance feedback to include the user in the retrieval loop. We can also mention the nice introduction proposed by Stork (2009) about the general topic of image processing for artwork analysis, where the author gave an overview of artistic characteristics and painting styles, side by side with the sophisticated image processing techniques proposed in literature to analyze them. These techniques can be used to recognize a particular artist by visual cues. See also the analysis of the relationships between painters based on their works by Bressan et al (2008), and the study on picture coding for paintings carried out by Graham (2009), highlighting the statistical regularities in artworks, in particular the use of efficient nonlinear tone mapping strategies.

Whichever set of features is chosen, especially in artistic content retrieval, the semantic gap limits the effectiveness of query by similarity, because of the heterogeneity of the visual appearance that different prototypes of a particular concept or artwork have. The use of *relevance feedback* strategies in information retrieval and in particular in CBIR systems is widely considered an effective way to mitigate this problem. In fact, the automatic association of low-level features to high-level semantics is still a very open problem, and the only practical way to identify what the user is looking for, is by including him in the retrieval loop with his positive or negative feedbacks.

The literature on this topic is (again) countless, since this problem can be faced from several points of view (computer vision, database management, human-computer interaction, artificial intelligence, and even psychology) (Zhou and Huang 2003; Crucianu et al 2004). However, four broad classes of relevance feedback techniques can be roughly identified:

- Query Point Movement approaches (QPM in short): they move the query point in order to create a better description of the user's need (Liu et al 2009).
- Feature Space Warping approaches (FSW in short): they manipulate the feature space in order to shape them in the direction of the users' feedbacks (Bang and Chen 2002; Nguyen et al 2007).
- Distance Metric Learning approaches (DML in short): they adjust the distance metric (so the notion of similarity between documents in the database) in order to adapt the comparison methodology to the users' feedbacks (Xing et al 2002)
- Machine learning approaches: SVM, boosting or transduction techniques are used to learn how to separate relevant samples from irrelevant ones (Tieu and Viola 2000; Tao et al 2006; Sahbi et al 2008; Grana et al 2011).

In a previous work, we employed in particular a transductive learning approach to solve the problem of relevance feedback (Grana et al 2011). The transductive process is a semi-supervised strategy particularly tailored for an interactive relevance feedback: in fact, we want to transfer labels from labeled samples to unlabeled ones, requiring the samples which are close in the feature space

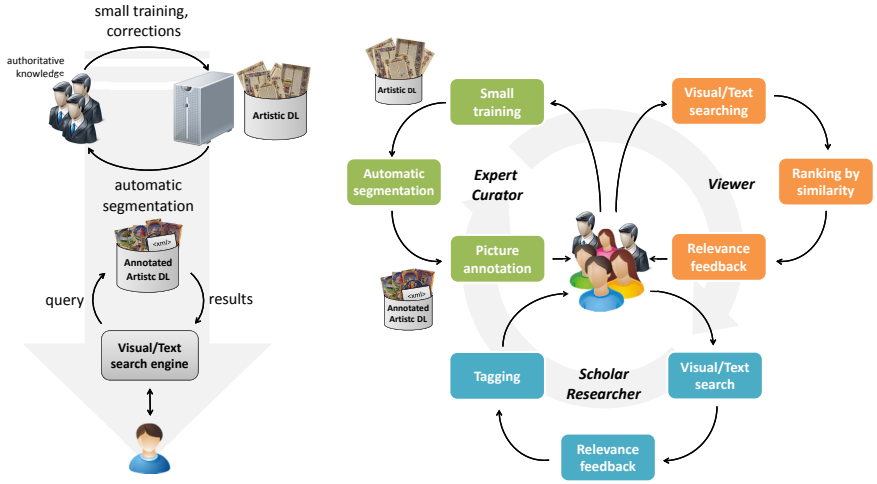


Fig. 1 In (a), we depict the standard approach used in information retrieval systems. An expert is required to include information of various nature into the system to allow the user to take advantage of proposed functionalities. In (b), we depict how things can change in a more flexible system like the one we are proposing. When the user is an expert, he can add his knowledge without the need of a structured representation (like tags or commentaries). When the user is a scholar or researcher, he can use the visual similarity and the relevance feedback to increase the amount of information of the system. When the user is a tourist, the system can be used as a simple information viewer.

to share the same label. In this scenario, the label assignment is provided by users through explicit feedback, constituting iteratively (and incrementally) the training set. A graph-based method can be used to solve the learning problem. However, considering the real-time and interactive constraints of our application, the usually higher computational requirements of learning-based strategies like this one, and the low amount of per-iteration feedbacks (which heavily impacts many learning-based strategies), in this paper we decided to focus on non-learning approaches.

3 Surfing and Tagging with a visual assistance

The majority of visual information retrieval systems follow the schema of Fig.1(a). Essentially, the system has a top-down design, and a professional effort (in terms of knowledge, documents and ontologies definition) is required to provide the user the full set of functionalities. Some image analysis and machine learning tools can be potentially exploited to facilitate the job, which nevertheless remains a professional prerogative. The annotated digital library is thus defined, often formalized as an ontology, and becomes as-it-is the center of the user experience. In this *content-centered* paradigm, the user does not

have a real role of intervention inside the structure: he turns out to be a simple viewer of the retrieval results, having no real interaction with the system.

In this paper, we want to suggest a structure more similar to the one in Fig.1(b). It is based on a *user-centered* paradigm, capable of putting together abilities, experiences and knowledge of different kinds of users, such as experts, art viewers, scholars and research communities. Instead of only assuming a static authoritative knowledge, often requiring long and laborious work of visual data annotation, we bet on visual similarity and relevance feedback to assist the process of knowledge addition by means of tags in a visual and interactive fashion.

In this context, as suggested by Datta et al (2008), a very interesting classification of multimedia systems can be proposed, based just on the user's intent:

- *Browsing*: when the user's end-goal is not clear, the *browser* performs a set of possibly unrelated searches by jumping across multiple topics;
- *Surfing*: when the end-goal is moderately clear, the *surfer* follows an exploratory path aimed at increasing the clarity of what is asked of the system;
- *Searching*: when the end-goal is very clear, the *searcher* submits a (typically short) set of specific queries, leading to the final results.

In our application, we tried to accommodate all these intents. The user can begin his analysis by *browsing* the pages of the documents, correcting the automatic segmentation or including a manual one if necessary. Whenever a particularly interesting detail is retrieved, the user can propose a tag to the selected picture and continue the exploration interactively, *surfing* by visual similarities. The system automatically answers with a set of similar pictures, which the user can further provide relevance feedback for. The results, marked by the user as similar, at the end may be given the same tags, so that the user will accomplish with minimal effort the otherwise demanding effort of tagging all pictures in the dataset when sharing the same visual content. Users' personal tags — after a validation task if deemed necessary to ensure reliability — become part of any other user' textual-based *searching* functionality, by providing a filtering on the visualized content very useful to focus the user attention on the particular section of the work he is mainly interested in.

4 Visual Search Engine

In this section, we detail the joint use of the covariance matrix for similarity retrieval and a new relevance feedback strategy with query remapping.

4.1 Covariance matrices

The motivations that brought use towards this solution stand in the nature of the images we are dealing with. As depicted in Fig.2, the miniature il-

Fig. 2 Some samples of miniature illustrations, in order to give an idea of the general properties of these images from the point of view of colors and gradients.

lustrations are very far for photorealism and very rough, despite being the result of an extremely precise hand work. Gradients are much more prominent compared with photos, with abrupt changes in intensity values over all color channels. Moreover the chromatic palette is limited, as it depends directly on the pigments and the manual effort with the available tools. Looking at those images, we immediately find out that the color is a fundamental component to take into consideration, and that textural information about gradients could help as well. Symbols, coats of arms, and crests usually have an even lower color palette and are surrounded by highly textured decorative parts. Natural scenes instead generally include a larger color palette. In addition, the location of color distribution can be extremely useful. Each coat of arms for example has a fixed color distribution in a fixed location of the miniature, while natural scenes or miniatures with animals in natural contexts have green on the bottom and blue on the top, which is radically different from the color distribution in indoor scenes. All these considerations confirmed us that the optimal feature for our context should have color and texture information, while accounting at the same time the location of these components inside the illustration itself. Many visual features considering separately these characteristics have been presented in Hurtut's survey. We hypothesize that an aggregate descriptor accounting all these information *and* the correlation between them would provide good results in our application. Moreover, given the nature of the miniatures and the generality we would like to include in our solution, we decided not to employ geometric constraints related to the layout of the page or any kind of semantic knowledge.

A possible solution for this kind of retrieval problem may be provided by the *covariance region descriptor*, originally proposed by Tuzel et al (2008) for region matching with applications to pedestrian and object detection, texture classification, and tracking. This approach allows to embed in a very compact

form a wide range of visual information (color, shape, spatial arrangement, gradients, etc.). Therefore, this global descriptor yields a straightforward solution for a low-dimensional feature representation: the matrix diagonal elements provide information on the variance of each source channel, while the off diagonal elements describe their correlation values.

Let I be a two-dimensional color image of width w and height h, $N = w \times h$, and

$$Z(x, y) = \phi(I, x, y) \tag{1}$$

be a function representing, for each pixel, the corresponding d dimensional feature vector, where ϕ can be any mapping of d simpler features, such as intensity, color, gradients, filter responses, etc. Let $\{\mathbf{z}_i\}_{i=1..N}$ be the set of N of these d-dimensional feature points inside Z. The image I is thus represented with the $d \times d$ covariance matrix of the feature points

$$\mathbf{C}_I = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{z}_i - \mu)(\mathbf{z}_i - \mu)^T \tag{2}$$

where μ is the mean of the feature points. The noise corrupting individual samples is largely filtered out with the average filter during covariance computation. This descriptor has a low dimensionality that, due to symmetry, can be further reduced to only $(d^2 + d)/2$ different values.

Particularly, for the image retrieval task, we mapped each pixel of the image into a 7-dimensional feature vector:

$$\phi(I, x, y) = \left[\frac{x}{w} \quad \frac{y}{h} \quad I_R \quad I_G \quad I_B \quad |I_x| \quad |I_y| \right]^T \tag{3}$$

thus, we employed normalized pixel locations, the RGB color values I_R, I_G, I_B, and the norm I_x and I_y of the first derivatives of the intensities, with respect to x and y, calculated through the filter $[-1 \quad 0 \quad 1]^T$. The resulting covariance of a region is a 7×7 matrix. The color information is carried by the color channel raw values, while textural information is accounted by gradients computed over x and y. Notice that the pixel locations variance (located in the first two diagonal entries of the matrix) is the same for all images with the same width to height ratio, but their covariances with the other features (located in the non diagonal entries of the matrix) is generally useful to correlate color and texture information with the different form factors of miniatures.

The covariance matrices do not form a vector space. For example, the space is not closed under multiplication with negative scalars. Most of the common machine learning algorithms, as well as relevance feedback approaches, assume that the data points form a vector space, therefore a suitable transformation is required prior to their use. In particular if we concentrate on nonsingular covariance matrices, we can observe that they are symmetric positive definite, and as such they can be formulated as a connected Riemannian manifold. In the following, matrices (points in the Riemannian manifold) will be denoted by bold uppercase letters, while vectors (points in the Euclidean space) by bold lowercase ones.

In order to rank images by visual similarity to a given query, we need to measure the distance between covariance matrices. Since covariance matrices do not lie on Euclidean space, we exploited the following distance measure for positive definite symmetric matrices as proposed by Förstner and Moonen (1999):

$$\rho(\mathbf{C}_1, \mathbf{C}_2) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_i(\mathbf{C}_1, \mathbf{C}_2)} \tag{4}$$

where $\{\lambda_i(\mathbf{C}_1, \mathbf{C}_2)\}_{i=1..d}$ are the generalized eigenvalues of $\mathbf{C}_1$ and $\mathbf{C}_2$.

Distance alone is not enough for our purposes. In fact most of the relevance feedback strategies (including ours) require to work on an Euclidean space in order to move the query and the other points with linear combinations. For the conversion from Riemannian manifold to Euclidean space, two steps are required (Tuzel et al 2008). The first step is the projection of the covariance matrices on an Euclidean space tangent to the Riemannian manifold on a specific tangency matrix $\mathbf{X}$. The second is the extraction of the orthonormal coordinates of the tangent vector.

The tangent vector of a covariance matrix $\mathbf{Y}$ is given by:

$$\mathbf{t}_{\mathbf{Y}} = \log_{\mathbf{X}}(\mathbf{Y}) = \mathbf{X}^{\frac{1}{2}} \log\left(\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}}\right)\mathbf{X}^{\frac{1}{2}} \tag{5}$$

where log is the ordinary matrix logarithm operator and $\log_{\mathbf{X}}$ is the manifold specific logarithm operator, dependent on the point $\mathbf{X}$ to which the projection hyperplane is tangent.

The orthonormal coordinates of the tangent vector $\mathbf{y}$ in the tangent space at point $\mathbf{X}$ are then given by the vector operator

$$\mathrm{vec}_{\mathbf{X}}(\mathbf{t}_{\mathbf{Y}}) = \mathrm{vec}_{\mathbf{I}}\left(\mathbf{X}^{-\frac{1}{2}}\mathbf{t}_{\mathbf{Y}}\mathbf{X}^{-\frac{1}{2}}\right) \tag{6}$$

where $\mathbf{I}$ is the identity matrix, while the vector operator at identity is defined as:

$$\mathrm{vec}_{\mathbf{I}}(\mathbf{Y}) = \left[y_{1,1}\ \sqrt{2}y_{1,2}\ \sqrt{2}y_{1,3} \dots y_{2,2}\ \sqrt{2}y_{2,3} \dots y_{d,d}\right] \tag{7}$$

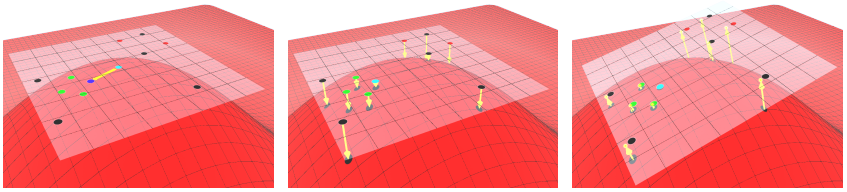
Substituting $\mathbf{t}_{\mathbf{Y}}$ from Eq. 5 in Eq. 6 we can write the simplified expression of the projection of $\mathbf{Y}$ on the hyperplane tangent to $\mathbf{X}$ as

$$\mathbf{y} = \mathrm{vec}_{\mathbf{I}}\left(\log\left(\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}}\right)\right) \tag{8}$$

In this way, after selecting an appropriate projection origin, every covariance matrix of size $d \times d = 49$ gets projected to a $(d^2 + d)/2 = 28$-dimensional feature vector on an Euclidean space.

Following a similar strategy, this process is invertible. We can compute the relative covariance matrix in the Riemannian manifold starting from the 28-dimensional feature vector lying on the Euclidean space using the following formulation:

$$\mathbf{Y} = \mathbf{X}^{\frac{1}{2}} \exp\left(\mathrm{vec}_{\mathbf{I}}^{-1}(\mathbf{y})\right)\mathbf{X}^{\frac{1}{2}} \tag{9}$$



(a) (b) (c)

Fig. 3 A qualitative visualization of the remapping process to the Euclidean space tangent on the new warping center. The cyan dot is the previous query point, while the blue dot is the current query point. Green and red dots represent the positive and negative feedbacks given by users. Black points are unlabeled samples. When the new query is computed by means of Eq.10, all the points are reprojected from the current Euclidean space (tangent on the previous query point) back to the original Riemannian manifold, and the new Euclidean space (tangent on the current query point) is thus computed. Then the FSW algorithm proceeds, by means of Eq.11.

4.2 Query Remapping Feature Space Warping

The relevance feedback strategy we proposed in this work, namely Query Remapping Feature Space Warping (QRFSW), aims at unifying the Query Point Movement (QPM) strategy and the Feature Space Warping (FSW), as already experimented by Chang et al (2009), but in a framework specifically conceived to efficiently deal with covariance matrices, which are not defined on Euclidean space.

Given a query point $\mathbf{q}$ in the feature vector space, k samples are retrieved by nearest neighbor search. By examining the results at each refinement step i, the user provides his feedback by specifying the relevance of M of these samples, forming two sets: $\{\mathbf{f}_p\}_i$ and $\{\mathbf{f}_n\}_i$, the relevant and irrelevant sets respectively. In the following, the consecutive movements of the query point $\mathbf{q}$ will be denoted as warping centers $\mathbf{w}_i$, with $\mathbf{w}_0 = q$.

QPM changes the position of the query point in the feature space according to the user's feedbacks, instead of leaving it in the original position $\mathbf{q}$. One of the most popular techniques for this purpose is the Rocchio's query movement formula (Rui et al 1997):

$$\mathbf{w}_i = \alpha \mathbf{w}_{i-1} + \beta \overline{\{\mathbf{f}_p\}_i} - \gamma \overline{\{\mathbf{f}_n\}_i} \tag{10}$$

where $\mathbf{w}_i$ is the new warping center after the shifting from the previous $\mathbf{w}_{i-1}$, and $\overline{\{\cdot\}}$ operator is the mean of a set. The set of parameters α, β and γ can be tuned to optimize the performance.

In the FSW, the feedbacks and the new warping center are used to warp the whole feature space, i.e. move all data samples $\{\mathbf{p}\}$ toward or away from the warping center $\mathbf{w}_i$. In particular, for each $\mathbf{p}_{i-1}$, its warped point $\mathbf{p}_i$ is given by

$$\mathbf{p}_i = \mathbf{p}_{i-1} + \lambda \sum_{j=1}^{M} u_j e^{-c|\mathbf{p}_{i-1} - \mathbf{f}_j|} \left(\mathbf{w}_i - \mathbf{p}_{i-1}\right) \tag{11}$$

where

$$u_j = \begin{cases} +1 \text{ if } \mathbf{f}_j \in \{f_p\} \\ -1 \text{ if } \mathbf{f}_j \in \{f_n\} \end{cases} \tag{12}$$

Two global coefficients c and λ are required to control the influence of each feedback to the samples.

The overall parametrization is very flexible, since it allows to easily switch between the two extreme algorithms: it becomes an example of a pure QPM by setting $\alpha = \gamma = \lambda = 0$ and $\beta = 1$, while it reverts to a pure FSW by setting $\alpha = 1$ and $\beta = \gamma = 0$. In our experiments, the whole set of parameters has been tuned with a grid search procedure on a portion of the dataset: the parameters were set varying within the $[0, 1]$ interval, with a step of 0.05.

Summarizing, when the user requires a refinement of a similarity search regarding a previously selected image, the new query point is being computed by linear combination of positive and negative feedbacks, then the FSW is applied to change the feature space accordingly, and finally the new results are re-ranked based on the newly warped distances.

As previously mentioned, many relevance feedback algorithms require a vector space to work properly. Both QPM and FSW fall in this category, since linear combinations are required to move the initial query to the new reformulated one and warp the whole set of feature points accordingly. Since covariance matrices lie over a Riemannian space, the MSFSW as detailed by Chang et al (2009) cannot be used as-it-is. A projection on the Euclidean space tangent at a point on the manifold becomes necessary, as suggested by Tuzel et al (2008) and reported here in Eq. 8. However, since this mapping is a homeomorphism around the neighborhood of the point, the structure of the manifold is only locally preserved: therefore the choice of the tangency point impacts directly on the quality of the projected vectors (in terms of optimal correspondence between the distances computed on the Riemannian manifold and those computed on the tangent space). For this reason, although a single mapping from Riemannian manifold to Euclidean space tangent at the initial query point works, the following space warps increase the inconsistencies between the underlying Riemannian structure (the real data) and the Euclidean projection. In other words, following the original MSFSW algorithm in our Riemannian context, each iteration step would lead to an increasingly bad mapping, resulting in a increasingly worse global performance of the relevance feedback itself.

To solve the problem, we propose to employ an intermediate step of reprojection (remapping) around the newly computed warping centers at each step i. QRFSW works iteratively according to the following sequence of steps:

1. Given the previous warping center $\mathbf{w}_{i-1}$ and feedbacks $\{\mathbf{f}\}_i$, the new warping center $\mathbf{w}_i$ is computed by means of Eq. 10 (Fig. 3(a));
2. All points $\{\mathbf{p}\}_{i-1}$ and $\mathbf{w}_i$ are reprojected on the manifold (Fig. 3(b)), defining the set of remapped points (matrices) $\{\mathbf{R}\}_i$, exploiting Eq. 9:

$$\mathbf{R}_i = \mathbf{W}_{i-1}^{\frac{1}{2}} \exp\left(\mathrm{vec}_{\mathbf{I}}^{-1}\left(\mathbf{p}_{i-1}\right)\right) \mathbf{W}_{i-1}^{\frac{1}{2}} \tag{13}$$

3. Now the tangent space at the new warping center in the Riemannian space $\mathbf{W}_i$ is taken into consideration: all remapped points on the manifold $\{\mathbf{R}\}_i$ are mapped into the new Euclidean space (Fig.3(c)), exploiting Eq. 8:

$$\mathbf{r}_i = \mathrm{vec}_{\mathbf{I}}\left(\log\left(\mathbf{W}_i^{-\frac{1}{2}}\mathbf{R}_i\mathbf{W}_i^{-\frac{1}{2}}\right)\right) \tag{14}$$

4. At this point, the FSW is applied on the set $\{\mathbf{r}\}_i$, as in Eq. 11, finally obtaining the new set of points $\{\mathbf{p}\}_i$.

Notice that on the first iteration only the feedbacks $\{\mathbf{f}\}_0$ must be initially mapped to the Euclidean space tangent at the query point. Thus only the new warping center $\mathbf{w}_1$ must be remapped, since the set of remapped points $\{\mathbf{R}\}_1$ would be equal to the original dataset itself.

Additionally, we observed that for our purposes the use of negative samples, as well as the presence of the previous warping center in Eq. 10, can have a negative impact on the algorithm's performance. This is especially true when the discriminative power of the employed feature for a particular class of images is low, or when the user selects some images as negative feedbacks because these contain semantically different concepts, despite the visual appearance being similar. This may cause some still undiscovered positive samples to be pushed away from the query. For this reason, we chose to compute the new warping center only as an average of the positive feedbacks collected so far. In this way we can further simplify the system getting rid of the three QPM-related parameters ($\alpha = \beta = \gamma = 0$).

As it will be discussed in Section 5, our proposal improves the relevance feedback performance with respect to the previous approaches, while using a simpler set of parameters and working natively on a non Euclidean space such as the Riemannian manifold. Its main weakness is the scalability, since the feature space warping and the remapping could be prohibitive with very large datasets and therefore becoming increasingly more complex with the increasing number of the required refinements. In this context, probably a (fast) learning-based approach could be preferable. Usually tree-based learning algorithms as Random Forest (Breiman 2001) are notoriously quite fast; online strategies such as online boosting (Grabner and Bischof 2006) or transductive SVM (Joachims 1999) can be alternatively exploited to incrementally update the classifier with new samples. However, in our experience, we observed on average a limited amount of feedback submitted by users and a low amount of relevance feedback iterations requested, which undermines significantly the potential of learning-based techniques. Anyway, given the aforementioned locality preservation of the mapping, it is unnecessary to take all the points of the feature space into consideration: in a very large dataset context, we can apply the relevance feedback to the first few thousands of nearest neighbor results. In particular, we verified that a couple of thousands of pictures can be easily processed on an average CPU. The problem of nearest neighbor search with very large datasets, which has not been faced in this work (given the limited amount of artistic pictures currently under our availability), has been

Fig. 4 Example of pictures grouped by class. (1) is identified with "nassa" and represents ancient fish trap (33 pictures) (2) is identified with "Fido" and represents a symbol of the Estense family (36 pictures) (3) is identified with "rosa" and represents a rose inside a ring with a diamond, which is the symbol of Duke Ercole I of Este (21 pictures) (4) is identified with "stemma" and is a symbol of Ercole I d'Este (21 pictures) (5) is identified with "scudo" and represent a symbol of the Rovigo county (19 pictures) (6) is identified with "bacinella" and represent a fountain with flames (36 pictures) (7) is identified with "eagle" and it is the symbol of the papacy (37 pictures) (8) is identified "crowd", representing all the scenes where a crowd of people is depicted (9) is "butterfly" (10) is identified with "praying" and represents all the scenes where people praying are depicted (11) is identified with "putto" and contains naked angels pictures, (12) is "portrait" (13) is identified with "game" and depicts a set of animals like deers.

recently studied in literature, and a lot of interesting techniques can be exploited in future developments (Torralba et al 2008; Weiss et al 2008; Jegou et al 2011).

5 Experimental results

We will report results on the digitalized pages of the Holy Bible of Borso d'Este, duke of Ferrara (Italy) from 1450 A.D. to 1471 A.D. It is one of the best Renaissance illuminated manuscripts in the world, whose original is held in the Biblioteca Estense Universitaria in Modena (Italy). It is composed by 640 pages, with two-column layered text in Gothic font, spaced out with some decorated drop caps, enclosing thousands of painted masterpieces surrounded by rich decorations. These pages have been digitalized at 10 Mpixels. Then an automatic procedure (Grana et al 2010) has been adopted to segment the miniature illustrations. The set of images obtained from the segmentation process has been manually refined to define the final dataset $\mathcal{D}$ of 2281 pictures, publicly available for scientific purposes. [1]

In order to propose an evaluation as fair as possible, we firstly defined an objective ground truth. In collaboration with a group of art experts, we per-

[1] Download the Bible dataset at `http://imagelab.ing.unimo.it/files/bible_dataset.zip`

formed a manual classification of $\mathcal{D}$ obtaining a subset $\mathcal{D}' \subset \mathcal{D}$ of 13 classes, characterized by a clear semantic meaning and a significant search relevance (see Fig.4). As a result, 41% of the original dataset (903 images) have been uniquely annotated into those classes, while the remaining pictures are considered as distractors, often with similar color, shape and texture distribution.

5.1 Features evaluation

The first metric adopted to evaluate the performance of the retrieval engine is the Mean Average Precision (mAP), one of the most widely used metrics to represent system effectiveness. Average precision for a single query is calculated by taking the mean of the precision scores obtained after each relevant document is retrieved. mAP is then computed as the mean of average precision scores over the whole set of queries. mAP is a popular metric, and has proved to be stable both across query set size and variations in relevance judgments (Turpin and Scholer 2006). In addition, as suggested for the TREC retrieval competitions, since recall and precision alone are set-based metrics (conveying respectively the ability of the system to present all relevant items and the ability of the system to present only relevant items), for ranked lists the plot of the precision against the recall is more significant. To facilitate computing average performance over a set of different topics (therefore different number of relevant items), individual topic precision values are interpolated to a set of standard recall levels i (0 to 1, incremented by 0.1). In particular, the interpolation rule is to use the maximum precision obtained for the topic for any actual recall level greater than or equal to i.

In order to propose a valuable comparison, a significant corpus of visual descriptors has been tested against covariance matrices in this specific dataset. In particular, we relied on the code and the implementation proposed by van de Sande et al (2010), employing the following descriptors:

- RGB histogram (*localCH*), i.e a combination of three 1D histograms based on the R, G, and B channels of the RGB color space;
- Transformed color histogram (*localTCH*): RGB histogram obtained by normalizing the pixel value distributions, achieving scale-invariance and shift-invariance with respect to light intensity;
- Color moments (*localCM*), using all generalized color moments up to the second degree and the first order, for a total 27-dimensional shift-invariant descriptor;
- SIFT descriptor, as proposed by Lowe (1999), which describes the local shape of a region using edge orientation histograms producing a 128-dimensional feature vector
- RGB-SIFT descriptor(*rgbSIFT*), computed for every RGB channel independently, for a total 3×128-dimensional feature vector;
- RG-SIFT descriptor(*rgSIFT*), computed for R and G channels independently, for a total 2×128-dimensional feature vector;

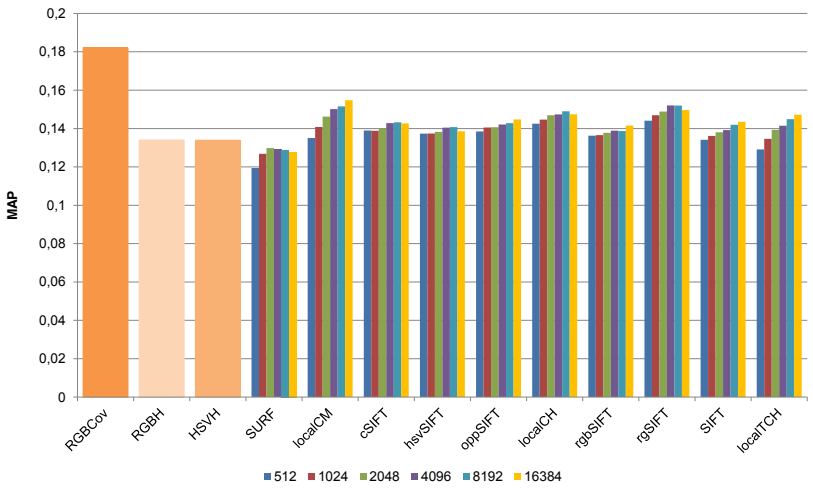


Fig. 5 Comparison of the mAP values obtained using different features. The first 4 in the chart are global, while the others are obtained as histograms of visual words, with dictionaries at increasing sizes. The covariance descriptor performance is presented in the leftmost bar.

- HSV-SIFT descriptor (*hsvSIFT*), computed converting the original image into the HSV color space, and considering each channel independently, for a total 3×128-dimensional feature vector;
- Opponent-SIFT descriptor (*oppSIFT*), describing all of the channels in the opponent color space (Tuytelaars and Mikolajczyk 2007) using SIFT descriptors;
- C-SIFT descriptor (*cSIFT*), as proposed by Burghouts and Geusebroek (2009), using the C-invariant color space which eliminates the remaining intensity information from the opponent channels;
- SURF, a scale- and rotation-invariant interest point detector and descriptor which uses integral images and other optimization and approximations to reduce the computational time.

All these descriptors were extracted using the Harris-Laplace region detector. A codebook has been created for every descriptor through a k-means clustering over 10% of the annotated dataset, randomly selected among all the classes in order to ensure an equal amount of visual information for each of them. The employed distance function is the histogram intersection. The sizes k of the codebooks have been determined empirically. In fact, since the clustering is a process of data compression, too small k's (large compression ratio) will force diverse keypoints into the same visual word reducing the quality of the representation; instead too large k's (small compression ratio) might lead to a sparse representation with similar keypoints mapped into different visual words, increasing the computational requirements without any real benefit. Therefore in our experiments we tested values of k between 2^9 and 2^{14}.

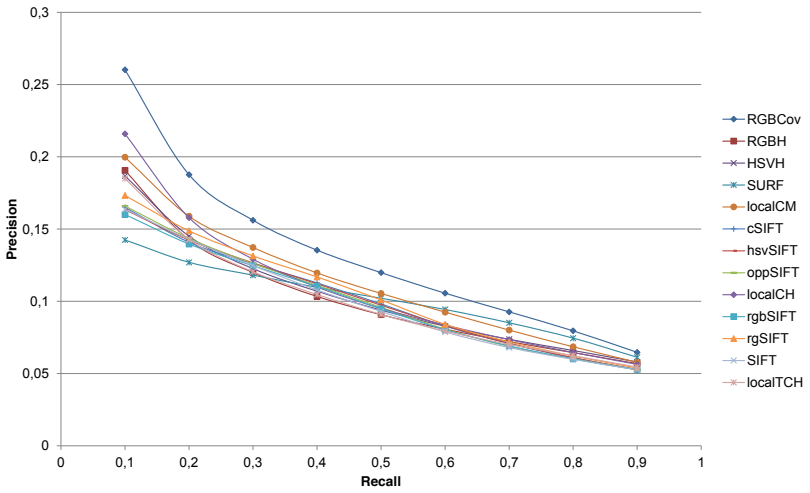


Fig. 6 Average precision for the proposed feature. Individual topic precision values are interpolated to a set of standard recall levels by interpolation. The covariance descriptor curve tops other features at every recall level.

Besides these local features, defined within the bag-of-words model, we considered three pure global features:

- RGB Histogram (*globalCH*), where each component is quantized into 8 values, resulting in a 512-bins histogram;
- HSV Histogram (*globalHSVH*), where each component of the HSV color model is quantized into 8 values, obtaining again a 512-bin histogram;
- Covariance matrix descriptor (*RGBCov*), which integrates color and gradient information, as presented in Section 4.1.

Fig.5 shows the performance comparison of the proposed features. The covariance matrix tops with 18.2%, while the RGB and HSV histograms largely fall behind. Surprisingly, also the bag-of-words histograms of local descriptors provide a significantly lower performance. Probably the interesting points extracted by the Harris-Laplace detector and the following k-means are biased over the inevitable amount of decorative content (especially when the illustration is not a rectangular region). The best performing local features are the rgsift (from 14.4% up to 15.2%) and the color moments (from 13.5% up to 15.2%), depending on the codebook size. As the chart shows, the increasing dictionary size generally does not help the retrieval in a substantial way, and certainly the increasing complexity of k-means and similarity retrieval does not match the benefits. In many cases, the transition from 4k codebooks to 8k and then to 16k codebooks lead to a worse (SURF, csift, hsvsift, rgb-histogram, rgsift) or comparable performance. In the other cases, the gain is quite marginal, and it must be carefully considered if it is worth the additional amount of processing required.

Precision over recall values (with 8k codebooks), presented in Fig.6, essentially confirm the good performance of covariance matrix compared to local descriptors. The precision outcome constantly remains above the other features for every fixed recall level.

Therefore, considering the results obtained and the dramatically lower computational demands of the global features (compared to the bag-of-words approach, requiring region detection, descriptor extraction, k-means clustering and histogram of visual words computation), we can state that a global approach based on covariance matrices for this kind of pictures is a very effective choice.

5.2 Relevance Feedback evaluation

In this Section, an evaluation of the proposed relevance feedback strategies is presented. The process has been performed automatically in order to avoid human errors. At each feedback iteration, all the pictures belonging to a particular class have been considered as positive feedbacks, as a human being typically would do in front of the system, while the others as negative ones. Therefore, at each iteration the number of feedbacks is variable, which is indeed a realistic scenario. The window of results proposed to the user at each iteration in order to be evaluated is fixed to 20: this value is arbitrary nevertheless consistent with the user interfaces available in common image search engines on the web.

Using the aforementioned 13 classes, we defined a total amount of 903 queries over the dataset, so multiple prototypes have been used as query for the same class. We chose a metric that matches the user's expectations in front of the application, that is the cumulative recall. It represents the overall recall provided by the system at each step $i = 1 \dots T$, and can be directly linked to retrieval quality for a user interested on collecting all pictures belonging to a particular class. The process automatically iterates the relevance feedback up to $T = 10$ refinements. This limit conveys the fact that the user will soon get bored and stops pursuing the search after T refinements at the most. While the first steps give an idea of the convergence capabilities of the algorithm, the last step gives an overall evaluation of the algorithm itself.

In order to show the effectiveness (and implicitly the usefulness) of relevance feedback in our context, we performed a comparison between the following algorithms:

- *Naïve relevance feedback* (actually no relevance feedback at all): the system discards the current set of n results and proposes to the user the next n, following the original rank given by the visual similarity;
- *QPM*: at each iteration, a new query is reformulated as the mean of positive feedbacks, that corresponds to $\alpha = 0.0$, $\beta = 1.0$, $\gamma = 0.0$;
- *FSW*: the Feature Space Warping is performed at each iteration (with $\lambda = 0.7$ and $c = 0.8$), without changing the query point (corresponding to $\alpha = 1.0$, $\beta = 0.0$, $\gamma = 0.0$);

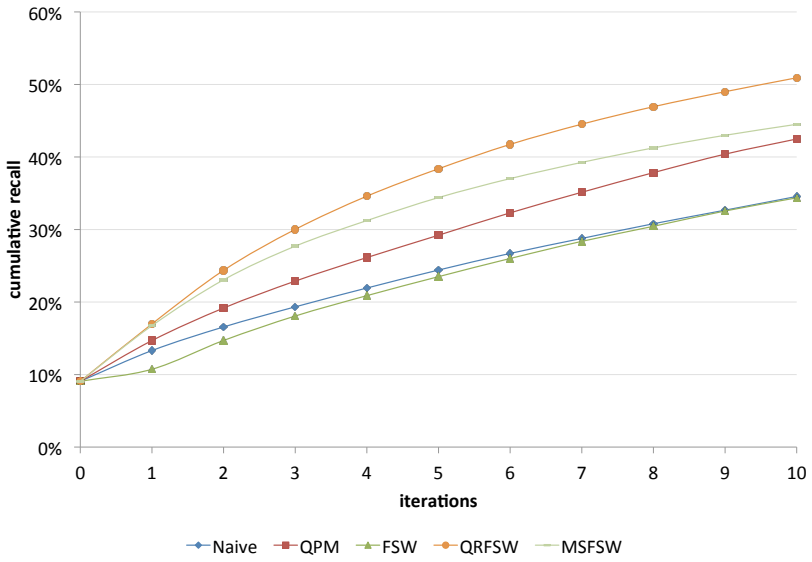


Fig. 7 Relevance feedback performance, measured in average recall at increasing refinement steps.

- *MSFSW*: the original Mean Shift Feature Space Warping proposal, which takes into account the influence of positive and negative feedbacks (where $\alpha = 0.2$, $\beta = 0.5$, $\gamma = 0.3$, $\lambda = 0.7$ and $c = 0.8$, optimized by grid search);
- *QRFSW*: our proposal exploiting the remapping of the entire feature space using the mean of positive feedbacks as tangent point for the conversion from Riemannian manifold to Euclidean space, with $\lambda = 0.7$ and $c = 0.8$.

The results are provided in Fig.7. In our discussion, we refer to iteration $i = 5$ as a mid-term checkpoint to evaluate the benefits of the relevance feedback after the initial refinement steps, and iteration $i = 10$ as a final checkpoint to draw a global evaluation of the algorithm itself.

The naïve technique is used as baseline for the comparison. It shows a progress in performance on the cumulative recall, up to 24.4% after the first 5 iterations and concluding with 34.6%. This straightforward approach does not actually help the user in the process of visual similarity search and tagging, and it should represent the upper bound in terms of necessary work to find and tag a specific class within the collection. The FSW used alone shows the worst performance, from 10.7% at the first iteration, 23.5% after 5 iteration and concluding with 34.4%. QPM, instead, constantly guarantees an increased amount of valuable pictures to the user, from 29.2% at iteration 5 up to to 42.5% after 10 iterations. The original MSFSW technique provides satisfying results in the initial part (36.8% after 5 iterations), with a flexion during the evolution of the search. The accumulation of inconsistencies between the original features in the Riemannian Manifold and how their mapped counterparts are evolved onto the Euclidean space tangent at the initial query point be-

comes a major source of performance loss. Moreover, the prevailing amount of negative samples induce the QPM and later the FSW to push away from the query a lot of good pictures yet to retrieve. The algorithm ends up with a cumulative recall of 48.8%.

On the contrary, QRFSW shows the steeper curve in the initial steps, up to 38.4% after 5 iterations, and the best per-iteration performance throughout the evolution of the query, finally concluding with 50.9%. The reasons why MSFSW loses in performance are basically the same reasons why QRFSW offers instead a better progression over iterations and the best final cumulative recall. By selecting at each iteration a new tangent point, we constantly keep under control the distortion of the mapped data: we therefore take into consideration at each iteration the best mapping available, in other words the Euclidean vectors more consistent with the covariance matrices describing the data. Moreover, by avoiding to weight the negative samples in the QPM part, QRFSW allows to focus solely on the good samples retrieved, and to arrange the feature space in such a way to favor the rising up of good pictures in the ranking.

6 User Interface Design

The user interface is a fundamental component of a multimedia system: it is the only part of the system which links directly to the user, and in an interactive scenario the amount and the quality of feedbacks have a direct impact on the algorithms beneath. For this reason, we developed an easy-to-use interface to enclose all the necessary functionalities; we also designed the system to be multitouch-friendly, so we included a set convenient multitouch gestures (like swiping, pinching, etc...) to trigger some basic functionalities like zooming, panning or scrolling. The upper part of the screen hosts a *preview bar* which facilitate the navigation through the manuscript. On the bottom, a *control bar* allows the selection (by tapping on the appropriate icon) of the main functionalities of the application: access to tags, zooming, manual editing of the automatic annotation and access to the miniature dashboard to manipulate the visual content. A global overview of the modules of the application, as well as screenshots of the user interface, are provided in Fig. 8 to describe the functionalities and the interaction workflow.

In the upper part of Fig. 8 we reported the preprocessing step described in a previous work (Grana et al 2010). The automatic segmentation provided by the system can be further refined manually (Fig.8a).

The preview bar on the top of the screen allows the user to scroll through the pages of the book. Once the page is selected, a higher resolution version is being presented to the user, complete with commentaries (Fig.8b). The details of the page can be observed using the zoom functionality (Fig.8c). To simplify the navigation, the user can also filter the visualized page previews by means of tags (Fig.8d): the textual search engine exploits both user-generated tags and meaningful keywords potentially provided by experts.

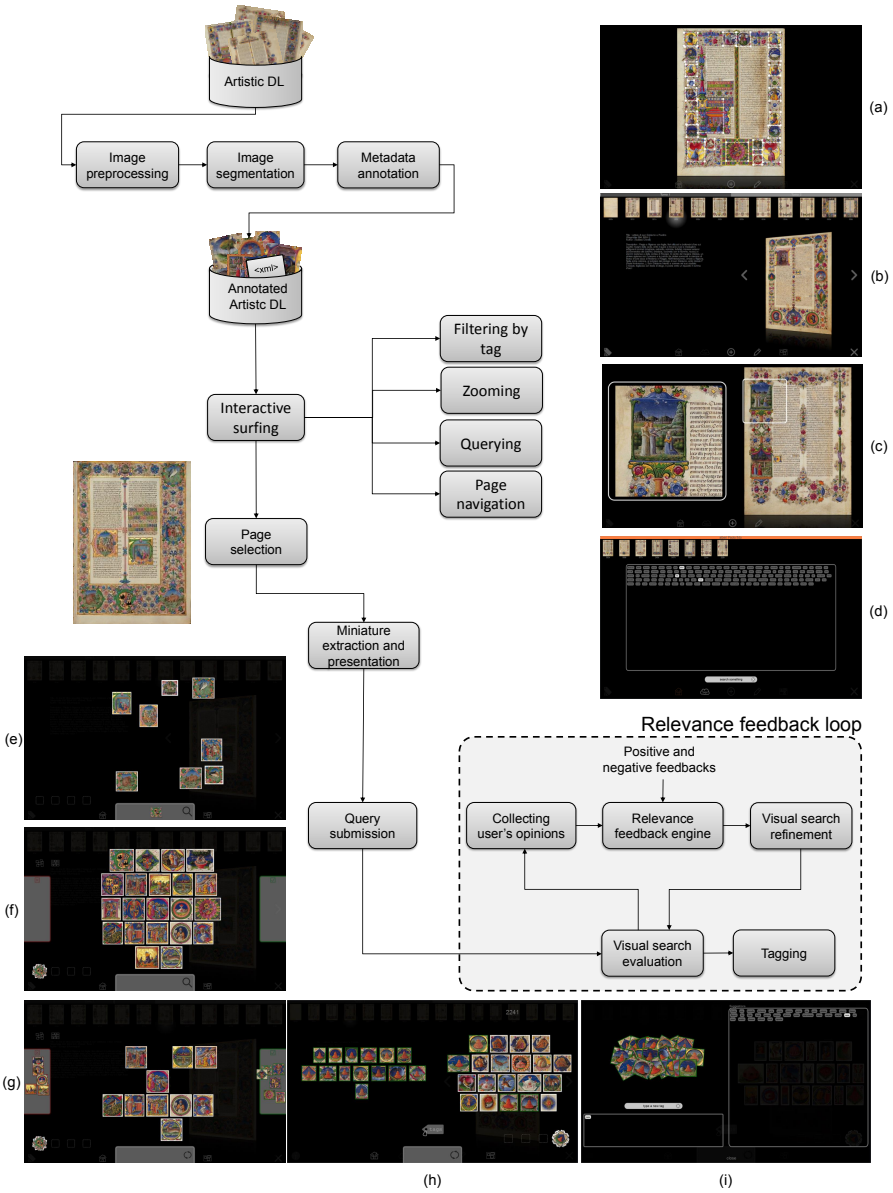


Fig. 8 Overview of the system, describing the functionalities and the interaction workflow. Screenshots of the application are provided as well.

The *miniature dashboard* is the core component that allows all extracted miniature illustrations to be freely managed by the user. Once a page is selected, the dashboard visualizes in overlay all the miniatures segmented from the current page. By dragging a query picture on the bottom box (Fig.8e), the user triggers the visual similarity search (Fig.8g). The system shows the user the most similar miniatures, employing either a grid-based layout or a random displacement, obtained by minimizing picture overlaps with a swarm-intelligence-based algorithm called cuckoo search (Yang and Deb 2009). At this point, the user can refine the visual search by suggesting the relevance of the obtained results. The feedbacks are collected into two areas (Fig.8g), depicted as green and red bordered boxes on the left and the right of the screen. The user is invited, according to his personal judgement, to fill these areas with relevant and not relevant miniatures by dragging them into. Using these feedbacks, the system tries to improve the current search providing more refined results. Once the search is completed and all the desired content has been hopefully found, a convenient user interface allows to assign a unique set of tags to all these pictures, conveying what we can call visually assisted tagging (Fig.8h,8h). These tags can be used to filter the pages, as previously showed in Fig. 8d.

7 Future works

The development directions that could be undertaken from this point are countless. In a engineering effort to extend this application to more illuminated manuscripts, it is easy to foresee the constitution of a smart library of illuminated manuscripts, in which the same similarity search engine could be used to perform visual searches among several books. This would provide to experts a valuable tool to explore and analyze in an integrated way new correlations between books, exploiting all the facilities of the digital paradigm. By refining the granularity of the segmentation, therefore moving from miniatures down to objects within miniatures, another powerful tool emerges. The possibility to search for single objects, in a context in which the symbolism permeates every detail of every page, could be extremely useful for researchers. Moreover, it is possible to design specific component with ad-hoc features to find out stylistic properties of the miniatures. These techniques could in other words be exploited to suggest the paternity of illustrations based on visual cues. Consider for example the stones on the grass, also visible in Fig.2, which are one of the stylistic signs of Taddeo Crivelli, the 15th century miniature painter who depicted the most of the illustrations of the dataset used in this paper. Another direction is the inclusion of textual and structural information coming from the rest of the page we excluded in our application. The possibility of linking automatically the text of the book, its translation and the corresponding commentary to each relevant object within the page is an interesting feature in sight of a complete system for the exploration of these masterpieces.

8 Conclusions

In this paper we presented a novel solution to explore interactively the miniature illustrations of illuminated manuscripts, suitable both for experts and general users. The proposal merges an appealing interface, a visual similarity based retrieval system and tag-based annotation system. The visual search relies on covariance information of color and gradients, a very compact yet effective descriptor in this context. An improved relevance feedback has been proposed, based on a remapping procedure in conjunction with a feature space warping, allowing to fasten and ease up the process of tagging by means of visual cues.

Considering the undoubted importance of Cultural Heritage in the contemporary society, we believe that this application may really improve the way illuminated manuscripts (and potentially similar artistic content) are approached. The possibility to cross the usual strict protection of these artistic masterpieces, opening them up to the public in their every intimate details, will be a fascinating and scientifically relevant direction for the next years.

References

Bang H, Chen T (2002) Feature space warping: an approach to relevance feedback. In: IEEE International Conference on Image Processing, pp 968–971

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Bressan M, Cifarelli C, Perronnin F (2008) An analysis of the relationship between painters based on their work. In: IEEE International Conference on Image Processing, pp 113–116

Burghouts GJ, Geusebroek JM (2009) Performance evaluation of local colour invariants. Computer Vision and Image Understanding 113:48–62

Chang Y, Kamataki K, Chen T (2009) Mean shift feature space warping for relevance feedback. In: IEEE International Conference on Image Processing, pp 1849–1852

Chen N, Blostein D (2007) A survey of document image classification: problem statement, classifier architecture and performance evaluation. International Journal of Document Analysis and Recognition 10(1):1–16

Corridoni JM, Del Bimbo A, Pala P (1998) Retrieval of paintings using effects induced by color features. In: Content-Based Access of Image and Video Databases, pp 2–11

Coustaty M, Pareti R, Vincent N, Ogier JM (2011) Towards historical document indexing: extraction of drop cap letters. International Journal of Document Analysis and Recognition 14(3):243–254

Crucianu M, Ferecatu M, Boujemaa N (2004) Relevance feedback for image retrieval: a short survey. In: State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction including Datamodels and Languages (DELOS2 Report

Dance CR, Csurka G, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision, pp 1–22

Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Computer Surveys 40(2):1–60

Förstner W, Moonen B (1999) A metric for covariance matrices. Tech. rep., Stuttgart University

Grabner H, Bischof H (2006) On-line boosting and vision. In: IEEE International Conference on Computer Vision and Pattern Recognition, vol 1, pp 260–267

Graham D (2009) Art statistics and visual processing: insights for picture coding. In: Picture Coding Symposium, pp 525–528

Grana C, Borghesani D, Cucchiara R (2010) Automatic segmentation of digitalized historical manuscripts. Multimedia Tools and Applications pp 1–24

Grana C, Borghesani D, Cucchiara R (2011) Relevance feedback strategies for artistic image collections tagging. In: ACM International Conference on Multimedia Retrieval

Hurtut T (2011) 2d artistic images analysis, a content-based survey. Tech. rep., Laboratoire d'Informatique PAris DEscartes - LIPADE - Université Paris Descartes

Jegou H, Douze M, Schmid C (2011) Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence 33:117–128

Joachims T (1999) Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning, pp 200–209

Lay JA, Guan L (2004) Retrieval for color artistry concepts. IEEE Transactions on Image Processing 13(3):326–339

Liu D, Hua K, Vu K, Yu N (2009) Fast query point movement techniques for large cbir systems. IEEE Transactions on Knowledge and Data Engineering 21(5):729–743

Lowe D (1999) Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, vol 2, pp 1150–1157

Luszczkiewicz M, Smolka B (2009) Application of bilateral filtering and gaussian mixture modeling for the retrieval of paintings. In: IEEE International Conference on Image Processing, pp 77–80

Marchenko Y, Chua TS, Jain R (2007) Ontology-based annotation of paintings using transductive inference framework. In: ACM International Conference on Multimedia, vol 4351, pp 13–23

Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. International Journal of Computer Vision 60(1):63–86

Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10):1615–1630

Nguyen G, Worring M, Smeulders A (2007) Interactive search by direct manipulation of dissimilarity space. IEEE Transactions on Multimedia 9(7):1404–1415

Rui Y, Huang T, Mehrotra S (1997) Content-based image retrieval with relevance feedback in mars. In: IEEE International Conference on Image Processing, vol 2, pp 815–818

Sahbi H, Etyngier P, Audibert JY, Keriven R (2008) Manifold learning using robust graph laplacian for interactive image search. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1582–1596

Stork DG (2009) Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature. In: International Conference on Computer Analysis of Images and Pattern, vol 5702, pp 9–24

Tao D, Tang X, Li X, Wu X (2006) Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(7):1088–1099

Tieu K, Viola P (2000) Boosting image retrieval. In: IEEE International Conference on Computer Vision and Pattern Recognition, vol 1, pp 228–235

Torralba A, Fergus R, Weiss Y (2008) Small codes and large image databases for recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp 1–8

Turpin A, Scholer F (2006) User performance versus precision measures for simple search tasks. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp 11–18

Tuytelaars T, Mikolajczyk K (2007) Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3):177–280

Tuzel O, Porikli F, Meer P (2008) Pedestrian Detection via Classification on Riemannian Manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10):1713–1727

Valle E, Cord M, Philipp-Foliguet S (2006) Content-based retrieval of images for cultural institutions using local descriptors. In: IEEE International Conference on Geometric Modeling and Imaging: New Trends, pp 177–182

Weiss Y, Torralba A, Fergus R (2008) Spectral hashing. In: Neural Information Processing Systems, pp 1753–1760

Xing EP, Ng AY, Jordan MI, Russell S (2002) Distance metric learning, with application to clustering with side-information. MIT Press pp 505–512

Yang XS, Deb S (2009) Cuckoo search via lvy flights. In: Second World Congress on Nature and Biologically Inspired Computing, IEEE, pp 210–214

Yen SH, Hsieh MH, Wang CJ, Lin HJ (2006) A content-based painting image retrieval system based on adaboost algorithm. In: IEEE International Conference on Systems, Man and Cybernetics, vol 3, pp 2407–2412

Zhou XS, Huang TS (2003) Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems 8(6):536–544

Zirnhelt S, Breckon T (2007) Artwork image retrieval using weighted colour and texture similarity. In: European Conference on Visual Media Production, p 1