Surfing on Artistic Documents with Visually Assisted Tagging

Daniele Borghesani, Costantino Grana, Rita Cucchiara
Dipartimento di Ingegneria dell'Informazione, University of Modena and Reggio Emilia
Via Vignolese 905/b, Modena, Italy
name.surname@unimore.it

ABSTRACT

This paper describes a complete architecture for the interactive exploration and annotation of artistic collections. In particular the focus is on Renaissance illuminated manuscripts, which typically contain thousands of pictures, used to comment or embellish the manuscript Gothic text. The final aim is to create a human centered multimedia application allowing the non practitioners to enjoy these masterpieces and expert users to share their knowledge. The system is composed by a modern user interface for browsing, surfing and querying, an automatic segmentation module, to ease the initial picture extraction task, and a similarity based retrieval engine, used to provide visually assisted tagging capabilities. A relevance feedback procedure is included to further refine the results. Experiments are reported regarding the adopted visual features based on covariance matrices and the Mean Shift Feature Space Warping relevance feedback. Finally some hints on the user interface for museum installations are discussed.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Algorithms

Keywords

Illuminated manuscripts, image retrieval, relevance feedback, covariance matrices, tagging, user interaction, visual similarity

1. INTRODUCTION

Among all the fields in which modern multimedia research can bring a leap forward in the user experience on Digital Libraries (DLs), cultural heritage is undoubtedly one of the most promising. All the plurality of masterpieces (paintings, books, manuscripts, even photos of sculptures and architectures eventually) can be effectively enclosed into a unique "paradigm" through digitalization, which allows a tremendous freedom of data elaboration: once we have the digital version of the artistic work, we can transmit it, preserve it, animate it, compare it with other works, search within it, and so on.

In cultural heritage, all these tasks produce a precious form of augmentation. The artistic work, which usually is enclosed or heavily protected due to its value and delicacy, now gains the capability to fill the gap with the public, to potentially show every intimate detail of itself. The advantages of an "augmented" artistic work are countless. Experts in the fields of art, religion and literature will have the possibility to deeply study all the details of the works, making them interoperable and ubiquitous through platforms and places, they could personalize their research, extend their research to a wider set of works with similar characteristics, even mix research results. On the other hand, normal people (museums visitors or people keen on art) can easily take a closer look to the work, increasing their experience and their interest.

Among the different forms of art, we concentrated our work on Renaissance illuminated manuscripts. Italy, in particular, has a significant collection of them, such as 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). These masterpieces contain thousands of valuable illustrations with different mythological and real animals, biblical episodes, court life illustrations, and some of them even testify the first attempts in exploring perspective for landscapes.

In this work, we present an ambitious project of a human-centered multimedia system for enjoying illuminated manuscripts DLs. We aim at defining a new paradigm for the interaction with the artistic documents. We propose a unified approach for efficient "surfing" of digital masterpieces both for experts and normal users, suitable also for attractive museum installations. The proposal employs state-of-the-art techniques for image segmentation, search and retrieval based on visual features and relevance feedback. In particular, we exploit covariance matrices as visual features, compared against other very common approaches such as Bag-of-keypoints or color histograms. We propose to adopt a relevance feedback methodology based on a query reformulation and a feature space warping to improve the retrieval results based on user feedback. Aside these techniques, the

MM'10, October 25–29, 2010, Firenze, Italy.

main novelty of the work is the joint use of visual similarity and relevance feedback for a *visually assisted tagging* of the manuscript, which eases the annotation process adding a remarkable amount of very useful textual information.

Until now, the usual approaches on domain specific DLs employ a great work of standardization under the woods, which aims at providing an ontology or at least an overall set of keywords in order to define and manage the metadata information. This *database-centered approach* is praiseworthy for archiving purposes, or for supporting expert users, but this structured knowledge representation could result too complex for people keen-on-art and their needs. Finally, the effort to design a specific ontology for each type of illuminated manuscript and the cost of manual annotation is often not affordable for the museum owning the masterpiece. Thus out proposal is the adoption of tagging, so successful in social media systems (e.g. Flickr and YouTube), and specifically visually assisted tagging which can exploit the existing commentaries and the interaction of users as a valuable alternative.

2. RELATED WORKS

Document analysis is one of the most explored fields in image analysis, and a plethora of work has been produced dealing with different aspects of the segmentation of the document. The seminal work of Nagy [18] provides an overview of the techniques proposed until some years ago for text segmentation, OCR and background removal. Chen et al. [4] provide a general partition of the classification approaches proposed so far. In particular the page can be classified using: image features (global and local descriptors for color, shape, texture, gradients and so on), physical layout features (a hierarchical description of the objects in the page, based on their geometric arrangement), logical structure features (a hierarchy of logical objects, based on the human-perceptible meaning of the document contents), textual features (the presence of keywords).

Several works tackle the physical and logical segmentation of the page, exploiting different rules on the page structure, such as geometric constraints over the layout. The majority of these works employ an XY-tree based representation, and graph or template matching approaches in order to perform classification. However, in our context, texture and color features may prove effective, since the patterns we are trying to classify have some distinguishing characteristics from both points of view. Texture features based on frequencies and orientations have been used in [11] to extract and compare elements of high semantic level, exploiting a block level page analysis. Nicolas et al. in [19] proposed a 2D conditional random field model to perform the same task. Histogram projection is used in [17] to distinguish text from images, while an approach based on thresholding, morphology and connected component analysis has been used in [13].

In the context of CBIR many other visual features have been proposed. Recently local features have attracted much interest, with SIFT [16] and SURF [2] being the most commonly adopted. Usually these descriptors are used with the bag-of-visual words approach: the set of centroids of the clustered training descriptors creates a so called *vocabulary*, then used to count the occurrences of descriptors belonging to every class within the objects of interest [5]. A recent work [26] discussed their use in CBIR: they perform similarly although SURF proved to be less time consuming.

Because of the complexity of the keypoint extraction and the importance of color in illuminated manuscripts, we explored the use of far simpler features which enclose color, edge and also spatial information, namely the Covariance Matrices, as proposed by Tuzel *et al.* in [25].

Until now, most of the activities on DL of illuminated manuscripts have been accomplished by manual annotation and indexing, but some interesting systems deserve to be mentioned. The AGORA [22] software performs a map of the foreground and the background and consequently propose a user interface to assist in the creation of an XML annotation of the page components. The Madonne system [20] is another French initiative to use document image analysis techniques for the purpose of preserving and exploiting cultural heritage documents. The DEBORA project [14] is another significant example of a complete system specifically designed for the analysis of Renaissance digital libraries.

3. SURFING ARTISTIC DOCUMENTS: A HUMAN-CENTERED APPROACH

Multimedia indexing and representation are tasks that are highly desirable to be automated with limited role of user interaction. In fact, the goal of most systems is to remove the user from the indexing loop and to achieve full automation. This is very important in light of the huge volumes of multimedia data. However, it is unlikely that fully automated multimedia archiving systems can be achieved in the near future. In order to achieve a usable multimedia system today, we need to involve the user in the retrieval loop. This is not just because of the lack of today's technologies to achieve a fully automated system, but mainly because different users have different interests in their multimedia data and, therefore, efficient, usable, multimedia representations need to be personalized [7]. The correct understanding of the user intent in the process of interaction with a multimedia system is fundamental for a successful design of the system itself. In [6], Datta *et al.* proposed a very interesting classification of multimedia systems based on the user intent, distinguishing three categories:

- Browsing: when the end-goal of the user is not clear; the *browser* performs a set of possibly unrelated searches, 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 he wants from 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.

These three modalities could require three different user interfaces, from a very general one in the first case to a very complex one in the last case, in order to satisfy the different degrees of expressive power needed. In our system we tried not to exclude any of these modalities, allowing powerful functionalities and complex queries for experts, but we designed the principal interface focusing on the surfing approach to artistic digital library, which we believe to be the typical behavior of people interested in exhibitions and art events. This result is accomplished leveraging on the use of

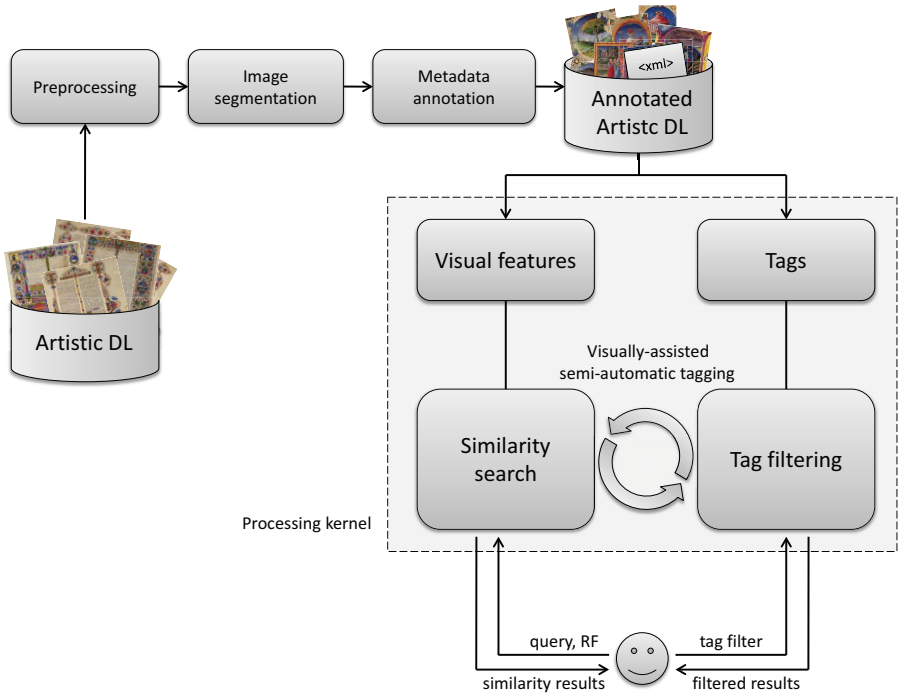

Figure 1: Architecture overview

both textual and visual information coming from tags and high level visual descriptors. The overall architecture of the system is proposed in Fig.1.

The process of annotating every picture of a DL is notoriously very boring and costly. This process can be extremely facilitated if the system, given a particular query, takes care about the extraction of the most similar pictures from the library to which the same annotation can be easily associated. This is the very basic idea behind the interaction between textual and visual information that we believe to be very useful in this context. This semi-automatic visually-assisted tagging procedure moves the user at the center of the multimedia experience. Starting from a clean system, with no prior information about the work (which can be trained with a small amount of ground truth - just some pages - to provide a first automatic segmentation of the more meaningful pictures), with no prior tags (except for the ones automatically extracted from experts-made commentaries and proposed by the system itself), the user begins his analysis by *browsing* the pages of the work, and correcting the automatic segmentation if necessary. Once the user finds a particularly interesting detail, he proposes a tag to the selected picture and then continues his analysis by *surfing* by visual similarity. The system automatically provides back a set of similar pictures, for which the user can further provide relevance feedback. The results marked by the user as similar at the end may be given the same tags, so with minimal effort the user will accomplish the otherwise demanding effort to tag all pictures in the dataset sharing the same visual content. Finally, the user can keep on analyzing the work by *searching* using specific and reliable (combinations of) tags, which will cause the system to filter out the visualized dataset allowing the user to focus his attention to the sections of the work he is mainly interested on. Basically, it is a virtuous loop in which the similarity search by visual content will allow the extraction of similar pictures (pictures which will likely share the same tags), and tags will help the user in the process of content search inside the manuscripts, and in the process of filtering results by topic.

Expanding this human-centered approach from a single artwork to a complete collection, a very powerful analysis tool emerges. The visual similarity can find meaningful results across different works, and an efficient use of tags to filter and organize documents, pictures and inner visual objects across different art works, allows the user to have literally the entire collection in his hands. That's the reason why we believe that this approach may help in the creation of a "smart library", capable to adapt to the user's needs using efficient but yet very simple user interaction approaches.

4. PICTURE SEGMENTATION

In order to relieve the user of the task of manual annotating the whole book collection, we employ a picture segmentation strategy described in [9]. Since the background has a nearly flat distribution and the overall chromatic range is quite distinguishing, we used automatic binarization using the Otsu algorithm (Fig. 2.a), which proved to be sufficiently robust to the noise. The connected components of the image are then labeled. The labeling in our context is particularly demanding since we work with high resolution images with tenths of thousands of connected components. To improve this stage, we employ a fast connected components analysis technique based on a 2×2 block optimization and an effective array-based data structure for label resolution [10].

Blobs smaller than a minimum area are removed in order to focus on the larger ones (Fig. 2.b). The contour of each blob is then followed and filled and the resulting pixels are used as a mask for the next stages of the processing (Fig. 2.d). This preprocessing stage is very important since it reduces the amount of pixels to analyze in the next stages roughly by 80%, leading to a dramatic reduction of the required processing times.

Once the preprocessing mask has been computed, the algorithm perform a block-based analysis with feature extraction, dimensionality reduction via embedding and at last a SVM classification, providing the segmentation of pictures inside the pages. Later on, the user provides some feedback on the segmentation (for example he manually annotates a new pictures previously undetected by the system or viceversa he deletes a wrong selection), the learning procedure starts again updating the segmentation on the remaining pages using the new positive (or negative) blocks as an update version of the training set.

5. VISUAL SIMILARITY USING COVARIANCE MATRICES

In order to accomplish an effective similarity retrieval upon these images, we relied on a simple yet effective feature which allows to consider both color and edge based information, that is covariance matrices. Computing the covariance region descriptor from multiple information sources yields a straightforward technique for a low-dimensional feature representation. A covariance matrix contains the variance of each source channel in its diagonal elements and the off diagonal elements describe the correlation values between the involved modalities.

The covariance matrices do not form a vector space. For example, the space is not closed under multiplication with negative scalers. 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.

Let I be a three-dimensional color image and F be the $W \times H \times d$ dimensional feature image extracted from I,

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

where the function ϕ can be any mapping such as intensity, color, gradients, filter responses, etc. Let $\{\mathbf{z}_i\}_{i=1..N}$ be the d-dimensional feature points inside F, with $N = W \times H$. The image I is 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 points. The noise corrupting individual samples is largely filtered out with the average filter during covariance computation. The descriptors are low-dimensional, and due to symmetry, C_I has only $(d^2 + d)/2$ different values.

For the image retrieval task, we use normalized pixel locations $(x/W, y/H)$, color (RGB) values and the norm of the first derivatives of the intensities with respect to x and y. Each pixel of the image is mapped to a seven-dimensional feature vector $\phi(I, x, y)$

$$\left[\frac{x}{W} \;\; \frac{y}{H} \;\; I_R \;\; I_G \;\; I_B \;\; |I_x| \;\; |I_y|\right]^T \tag{3}$$

where I_R, I_G, I_B are the RGB color values, and I_x, I_y are the intensity derivatives, calculated through the filter $[-1 \;\; 0 \;\; 1]^T$. The covariance of a region is a 7×7 matrix. Although the variance of pixel locations is the same for all images with the same width to height ratio, they are still important since their correlation with the other features are used at the nondiagonal entries of the covariance matrix.

In order to rank images by visual similarity to a given query, we need to measure the distance between covariance matrices. As already mentioned, the covariance matrices do not lie on Euclidean space, thus in [8] the following distance measure for positive definite symmetric matrices is proposed:

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

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

Unfortunately distance alone is not enough for our purposes. In fact to enable the user to provide relevance feedbacks, we need to work on an Euclidean space, which allows us to move the query and the other points with linear combinations. To this aim two steps are required[25]: the projection on the tangent space, and the extraction of the orthonormal coordinates of the tangent vector. The tangent vector of $\mathbf{Y}$ 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 to which the projection hyperplane is tangent.

The orthonormal coordinates of the tangent vector $\mathbf{y}$ in the tangent space at point $\mathbf{X}$ is 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 I is the identity matrix, and the vector operator at identity is defined as

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

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 gets projected to a 28-dimensional feature vector on an Euclidean space.

6. RELEVANCE FEEDBACK

While exploring an artistic collection, the user is interested in obtaining images similar to some of the results already obtained by tag searching or page browsing, possibly retrieving as many relevant results as possible. Unfortunately the semantic gap may limit the effectiveness of a

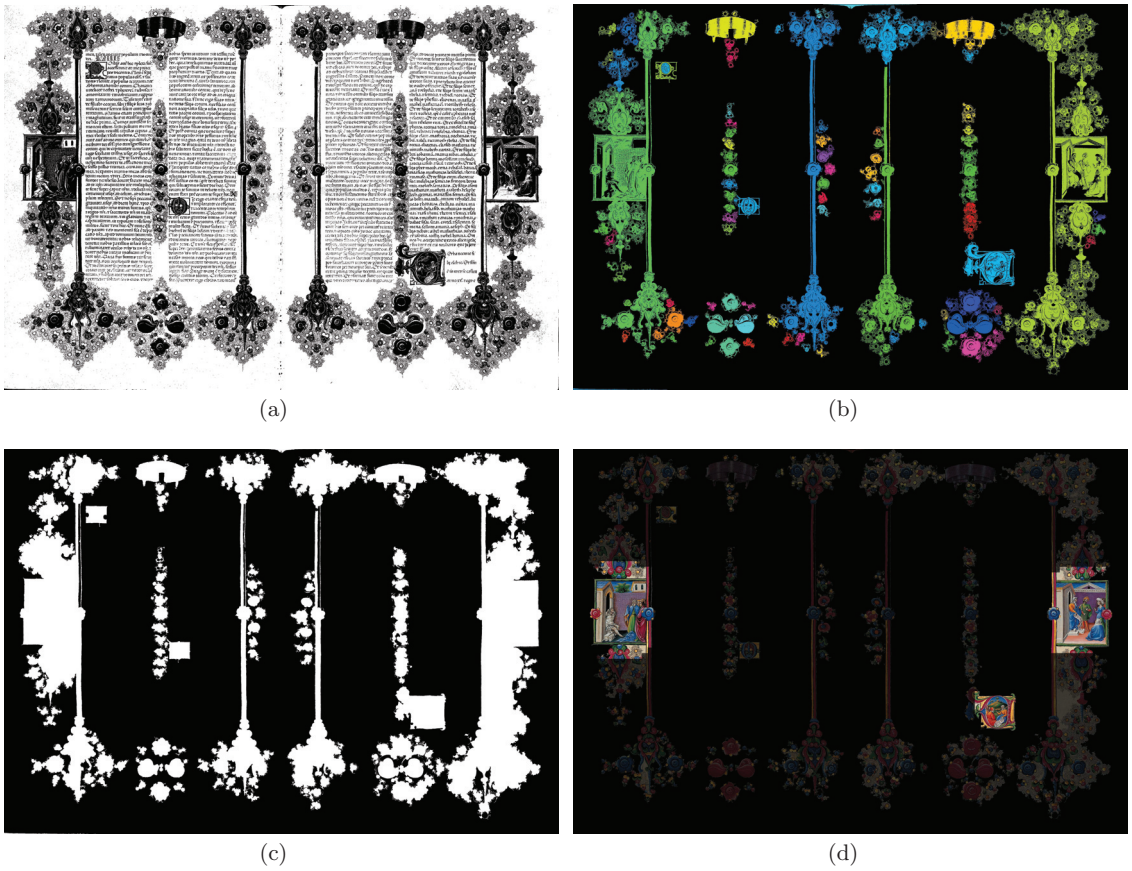

Figure 2: Steps of the segmentation stage. After the Otsu thresholding (a), connected components analysis is performed (b) followed by a filling procedure (c). Finally an SVM classification is adopted to extract meaningful pictures (d).

query by similarity, given the heterogeneity of the visual appearance of some prototypes of a particular concept (e.g. animal, plant, etc...). Relevance feedback gives the retrieval system a chance to improve its results by exploiting the extra information provided by the user through more elaborate techniques.

The proposals described in literature for relevance feedback can be roughly divided into three categories:

- Query reformulation: the user feedbacks are used to produce a refined query, trying to increase the likelihood of being near positive results and farther from negative ones. A typical example is the well known Rocchio's formula [23].
- Feature space transformation: manipulating the feature space by adjusting the weights of each feature component and move every sample or by modifying the similarity measure based on user's feedbacks;
- Online classifier learning: a classifier (e.g. SVM or an Adaboost) is immediately trained to separate relevant samples from irrelevant ones, and used to provide the new ranked results list.

In particular, the first two approaches have a quite different behavior depending on the data distributions properties, but generally they are the best performing techniques in the literature. In our application, we decided to apply a recent proposal [3], which combines the Feature Space Warping (FSW) approach [1] with the classic Query Point Movement (QPM), called Mean Shift Feature Space Warping (MSFSW). Given a query point $\mathbf{q}$ in the feature vector space, k samples are retrieved by nearest neighbor search. By examining the results, the user provides his feedback by specifying the relevance of M of these samples, forming two sets: $\{\mathbf{f}_p\}$ and $\{\mathbf{f}_n\}$, the relevant and irrelevant sets respectively. These are employed to move all data samples $\{\mathbf{p}\}$ toward or away from the warping center $\mathbf{w}$. In particular, for each $\mathbf{p}$, its warped point $\mathbf{p}'$ is given by

$$\mathbf{p}' = \mathbf{p} + \lambda \sum_{j=1}^{M} u_j \exp\left(-c|\mathbf{p} - \mathbf{f}_j|\right)(\mathbf{w} - \mathbf{p}) \qquad (9)$$

where the scalar value u_j is set to $+1$ if $\mathbf{f}_j \in \{\mathbf{f}_p\}$, and -1 if $\mathbf{f}_j \in \{\mathbf{f}_n\}$. Two global coefficients c and λ are required to control the influence of each feedback to each sample and the maximum moving factor of any point $\mathbf{p}$ toward or away from the warping center $\mathbf{w}$.

The original FSW algorithm fixes the warping center $\mathbf{w}$

Figure 3: Example of pictures grouped by class

to $\mathbf{q}$. Thus, the query point will always stay in its original position. Other points will move toward or far away from $\mathbf{q}$ based on its proximity to relevant and irrelevant sets. But, according to the analysis proposed in [3], FSW algorithm tends to perform poorly under Gaussian distributions when the query point is far away from the cluster center. For this reason, in the MSFSW, authors proposed to move the warping center instead of staying at $\mathbf{q}$. They suggest to adopt the Rocchio's query movement formula:

$$\mathbf{w}' = \alpha\mathbf{w} + \beta\overline{\mathbf{f}_p} - \gamma\overline{\mathbf{f}_n} \qquad (10)$$

where $\mathbf{w}$ is the warping center (initially set to $\mathbf{q}$), $\overline{\mathbf{f}_p}$ and $\overline{\mathbf{f}_n}$ are the mean of the set $\{\mathbf{f}_p\}$ and $\{\mathbf{f}_n\}$. Another set of parameters α,β and γ is required, and must be tuned to optimize the performance.

With the above formulations, the MSFSW algorithm provides a flexible parameterization for switching between the two extreme algorithms: QPM by setting $\alpha = \gamma = \lambda = 0$ and $\beta = 1$, and FSW by setting $\alpha = 1$ and $\beta = \gamma = 0$. Given the final user target of our application, exposing the parameters configuration to the user was out of question. Thus, we determined the parameters configuration which provided best results on a small initial training set, using an automatic exhaustive search procedure.

From the above equations, it is clear that we need a way to compute a linear combination of the feature vectors. For this reason, we employed the projection of the covariance matrices on the tangent space previously described. As mentioned before, the projection requires a point from which determine the orthonormal coordinates of the tangent vector (i.e. the vector in the Euclidean space). Our experiments confirm that the choice of this point is fundamental to guarantee an optimal correspondence between the distances computed on the Riemannian manifold and those computed on the tangent space.

Thus, when the user requires a refinement of a similarity search of a previously selected image, we project the whole feature space on the chosen query point (i.e. the covariance matrix of the selected image). Then we rank the results and perform further refinements based on subsequent relevance feedbacks on this specific projection. We could also perform the initial similarity search on this Euclidean space, but this procedure is computationally expensive, so for this purpose we rely on Eq. 4 which computes the distances directly within the Riemannian manifold.

Feature	MAP
RGB Histogram (w/o weighting)	0.467
RGB Histogram	0.525
Bag of Visual Words (SURF)	0.519
Covariance Matrices	0.647
Covariance Matrices w/ RF(10)	0.750

Table 1: MAP values for different features

7. EXPERIMENTAL RESULTS

Our proposal is independent by the specific illuminated manuscript, since only some sampled pages were exploited to tune the parameters. To discuss some experiments, here we report on the results on the digitalized pages of the Holy Bible of Borso d'Este, duke of Ferrara (Italy) from 1450 to 1471 A.C., which is considered one of the best Renaissance illuminated manuscript in the world. Tests has been performed among a dataset of 640 high resolution digitalized pages (1947x2792). These have been automatically segmented but results have been manually refined to provide a precise ground truth and half or them are used for training and half for testing. Each page of the dataset is an illuminated manuscript composed by a two-column layered text in Gothic font, spaced out with some decorated drop caps. The entire surrounding is highly decorated. Using the procedure described in [9], followed by a manual correction of the extracted pictures, we obtained a dataset of 2667 pictures.

We manually annotated a subset of 500 pictures subdivided in a set 40 of clearly distinguishable classes (e.g. rose, little angel, deer, portrait, group of people, person, boar, fountain, bird, butterfly, etc... See Fig. 3). Note that we explicitly avoided to include concept without objective appearance properties (i.e. Jacob, Abram or other biblical characters depicted). In fact the ability to distinguish such visual content can be provided only exploiting tags or textual descriptions.

The metric adopted to evaluate the performance of the retrieval engine and the improvement provided by the relevance feedback is the Mean Average Precision (MAP), one of the most widely used system 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, with relevant documents that are not retrieved receiving a precision score of zero. MAP is then computed as the mean of average precision scores over the whole set of queries. MAP is a popular metric, and has been shown to be stable both across query set size and variations in relevance judgments [24].

Four different features were compared: as baseline, we selected the plain RGB Histogram, where each component is quantized to 8 values, resulting in a 512-bin histogram. A first observation of the results showed that the decorated frames were attracting similarity more than the real picture content, thus every pixel contribution was weighted using a two dimensional Gaussian kernel. Using this weighting function, the RGB Histogram was computing, showing a significant improvement in the results. To verify the importance of gradient based features, the Bag of Visual Words

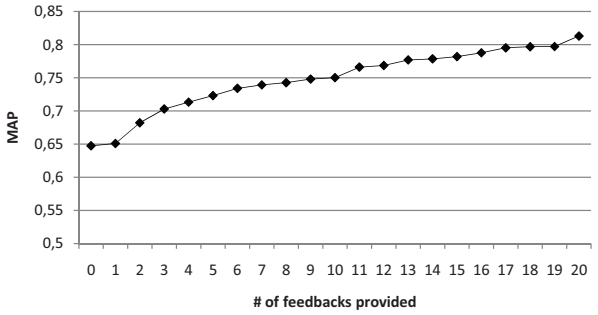

Figure 4: MAP values obtained changing the number of feedbacks provided

approach was selected, using SURF as local features. The dictionary was fixed to 1000 visual words and obtained using the k-means algorithm on the training set. The last feature used is the covariance feature described in the previous sections. Both the Bag of Visual Words and the Covariance Matrices used the Gaussian weighting function.

Table 1 shows the performance of the four features: in our dataset, the best performances were obtained by the Covariance Matrices. The RGB Histogram has interesting performance, even superior to the Bag of Visual Words, which is definitely suffering the lack of color information.

The next experiment deals with the effect of relevance feedback in this context. Since the users task is mainly exploratory, we believe that it is unreasonable to expect the users to refine their queries with more than one iteration of relevance feedback. For this reason we concentrated on a single refinement action, testing the MAP improvement with respect to the number of feedbacks provided. The MSFSW parameters were set to $\alpha = 0.2$, $\beta = 0.5$, $\gamma = 0.3$, $\lambda = 0.7$, $c = 0.8$. The feedbacks were given by labeling as relevant or irrelevant the ranked results in the order with which they were presented by the system (the most similar result was judged first and so on).

Fig. 4 shows the MAP value at different levels of provided feedbacks. The interface allows the user to inspect 10 to 15 results after the similarity search, then with a single touch or click the user marks which results are relevant and the system assumes the others are irrelevant. With this approach the MAP should improve from the 0.65 of the original query to a value between 0.75 and 0.78, depending on the number of feedbacks provided. An example of query refinement is provided in Fig. 6(b), where the "fish trap" is selected, and the relevance feedback results are shown in Fig. 6(c).

8. EXPLOITING TAGS FOR RETRIEVAL

In all art works, textual information has a fundamental importance. In particular, we refer to all the additional information that the experts have made available, such as, in the case of the Holy Bible of Borso d'Este, the identity of a particular person, a technical explanation about a symbol, a feeling, a connection between a scene and a part of the whole story and so on. All these data are added using commentaries and the work of experts.

The mechanism of tag management is organized in a hierarchical fashion:

- Document tag: this tag is typically the name of the work, and it corresponds to the selection of a specific document inside the library.
- Page tags: these tags are automatically extracted from the document commentary, using a process similar to the tag cloud algorithm. The rationale behind it is that the words used by an expert to describe (even quite generally) a page must correspond to the actual content of the page, so searching these tags will return most of the results the user is interested in, together with many other unrelated results.
- Picture tags: each picture, automatically extracted by the system or corrected by the user, can be described using one or more tags. Sometimes the tags are necessary to retrieve the exact content, especially when the content is difficult to interpret (such a symbol), difficult to recognize (due to the manual painting process) or even difficult to link to a known object or person.
- Object tags: inside each picture there is a world of objects. In some cases, these objects are very important for experts (for example the white stones in the ground, which experts recognized as a sort of signature of the painter). For this reason, a deep granularity of tags down to object level is a valuable addition to the system.

In this work we do not propose to fuse visual and textual information at query level. Many works, propose a fusion of multi-modalities as weighting differently text and visual features [12], adopting reranking mechanisms [21] or creating a visual tag dictionary [27]. Here we suppose to have reliable textual information, although with a low precision degree, since the tags referred to a whole pages cannot be precisely associated with single parts of the page itself. Thus we propose to use tags to filter out results of the retrieved-by content engine or to exploit tags for initial queries, than manually refined by visual search.

The selection of tags employs an AND logic, which selects a subtree of the hierarchy. For example, selecting a document tag focuses the attention of the system inside that document; selecting a document and a page tag focuses the attention of the system inside that particular page, visualizing all the pictures extracted from that page; finally, selecting a document and a picture tag will visualize all the picture of the document with that tag. The defined tag navigator module is provided in the user interface as discussed in the next section. Document and page tags are assigned a priori, in particular a tag cloud-like approach is used and accordingly customized to extract the most relevant keywords from the commentary. Picture tags instead are built incrementally by users.

Tags, especially in social contexts, have the well known problem of consistency: the information submitted by users is "subjective", it could not be consistent among different users, and sometimes even within the user itself, since the social tagging process is not standardized. In literature, a lot of works deal just with this kind of problem [15]. In case of domain-specific applications like artistic documents, this could not represent a problem, if tags are provided by expert peoples only in an automatic manner, exploiting commentaries, or in a semi-automatic way visually assisted by content-based search. In this first version of the project,

Figure 5: User interface and tagging screenshots

in fact, our system foresees two kinds of users: experts with high privileges which will have the possibility to edit textual information and create new tags, and normal users with no privileges which will enjoy the results. Then we will provide infrastructure to share "subjective" tags too, with their intrinsic degree of uncertainly, with their expressive power related with affective experiences.

9. UI DESIGN FOR A HUMAN-CENTERED TOOL

The user interface and the user interaction paradigm is a fundamental aspect for a multimedia system, because it is the only part of the system which will link directly to the user's emotion. For this reason, if the proposed user interaction is good, the user will be pleased to come back using the application.

For our system, we were inspired by the most successful user interfaces proposed in the market and in literature, and we came out with a very easy-to-use, minimal but yet effective interface which enclosed all the necessary functionalities. The entire interface is multitouch-friendly, since the finger is the most easy-to-use pointing device we have. The most complex functionalities are described by (a small amount of) clearly understandable icons, while —when it is possible— all the other minor functionalities (for example selection, navigating, zooming and so no) are generally triggered by convenient gestures (like swipe, pinch, etc...). Screenshots of the UI are provided in Fig. 5 and 6.

In the next few sections, we include a brief discussion about the software modules which provide the main functionalities of the UI. These modules are indeed strongly representative of the usability principles described in this paper:

- A natural user interface, which proposes a touch-based interaction (even for the complex tasks) and an intuitive approach to the required functionalities (in terms of icons and gestures)
- An emotional interaction, with a particular care —in accordance with the art works beauty— on the aesthetic of the interface elements and the interactive animations
- An efficient way to propose both the traditional functionalities (for example the categorization of pictures, addition of metadata to pictures) side by side with the innovative similarity search, proposing a unique mix of semi-automatic visually assisted tagging and advanced retrieval by visual content
- A social interface, very useful to people that want to share opinions about the art work and want to collaborate in order to improve the knowledge about the work itself
- A strong human-centered approach, which selects and implements the underneath technologies in such a way to be built around the needs of the type of users that will be using the tool and improving results by relevance feedbacks

(a) (b) (c) (d)

Figure 6: Picture level and similarity search screenshots

The entire surfing capabilities of the system is built around an overlay component called *tag navigator* which provides the user the access to the tags available in the system (Fig. 5(c)), allowing to move to higher or lower hierarchy levels. When selecting one or more tags, the system filters results in background, in order to provide the exact subset of data on which the user wants to focus on. We can exploit tags to conveniently limit the amount of information visualized to the user: by selecting more tags, we can provide an AND logic for a more fine grained filtering capabilities.

To each page of the document, the system can provide an automatic annotation, in the form of a list of selections (Fig. 6(b)). Each selection corresponds to a region of interest of the page, to which we can assign one or more tags (Fig. 5(d)). Once a particular page is selected, an *annotation editor* can be invoked in order to correct the automatic annotation provided by the system, or providing a custom annotation if necessary (Fig. 6(a)). A zooming functionality is included to perform a fine grain selection of pictures (Fig. 5(b)).

The entire process of managing tags for picture, assign a picture to folder of favorites pictures, or add comments to pictures is performed within a component called *social dashboard*. This component is currently highly experimental: it will be completed with the required functionalities as we start collecting significant amounts of user opinions. The set of pictures subdivided by tags, as well as the favorites set of pictures customized by users, will be visualized as a stack of pictures in an appropriate view called *pile viewer*. Comments are saved in this dashboard in the form as a sticky note (to mimic the traditional way librarian use to take notes about art works).

The entire process of visually assisted tagging is performed within the *similarity dashboard*. It is a fundamental component for the user experience in this application, since it cares about the visualization of the similarity search given a selected image (query) and gathers feedback information exploiting the pictures marked by users (through touching) as relevant (Fig. 6(c) and 6(d)). The similarity dashboard can subsequently trigger a refinement of the search exploiting the relevance feedback collected at each step, and finally can trigger the process of extending tags from query to relevant results, in order to improve the textual information available to the system and used to populate accordingly the tag navigator.

10. CONCLUSIONS

In this paper we presented an innovative complete solution which merges a modern interface, a visual similarity based retrieval system and tag based annotation to provide a unique user experience for the exploration and annotation of illuminated manuscripts. We believe that this application may really improve the way people approach this kind of artistic collection, exporting their beauty from old dusty books segregated in archives, to new fascinating interactive experiences.

In order to accomplish this goal, the next development will deal with further application improvements in terms

of usability and performance. To this aim efficient indexing techniques for high-dimensional data, that still lack in many commercial RDBMS, must be employed to significantly lower the system response-time. Moreover, expanding the audience by means of a web based version of the application could be a challenging but rewarding step. Of course this would require to also deal with problems of high quality content delivery, copyright and data protection, and correct content attribution policies.

11. ACKNOWLEDGMENTS

This work is supported by Franco Cosimo Panini Spa. We would like to thank Biblioteca Estense Universitaria for the availability of their digital library of illuminated manuscripts.

12. REFERENCES

[1] H. Bang and T. Chen. Feature space warping: an approach to relevance feedback. In *IEEE International Conference on Image Processing*, pages 968–971, 2002.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[3] Y. Chang, K. Kamataki, and T. Chen. Mean shift feature space warping for relevance feedback. In *IEEE International Conference on Image Processing*, pages 1849–1852, 2009.

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

[5] C. R. Dance, G. Csurka, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004.

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

[7] A. Elgammal. Human-centered multimedia: representations and challenges. In *ACM international workshop on Human-centered multimedia*, pages 11–18, 2006.

[8] W. Förstner and B. Moonen. A metric for covariance matrices. Technical report, Stuttgart University, 1999.

[9] C. Grana, D. Borghesani, and R. Cucchiara. Picture Extraction from Digitized Historical Manuscripts. In *ACM International Conference on Image and Video Retrieval*, July 2009.

[10] C. Grana, D. Borghesani, and R. Cucchiara. Optimized Block-based Connected Components Labeling with Decision Trees. *IEEE Transactions on Image Processing*, 19(6), June 2010.

[11] N. Journet, J. Ramel, R. Mullot, and V. Eglin. Document image characterization using a multiresolution analysis of the texture: application to old documents. *International Journal of Document Analysis and Recognition*, 11(1):9–18, 2008.

[12] L. Kennedy, S. Chang, and A. Natsev. Query-Adaptive Fusion for Multimodal Search. *Proceedings of the IEEE*, 96(4):567–588, 2008.

[13] A. Kitamoto, M. Onishi, T. Ikezaki, D. Deuff, E. Meyer, S. Sato, T. Muramatsu, R. Kamida, T. Yamamoto, and K. Ono. Digital Bleaching and Content Extraction for the Digital Archive of Rare Books. In *International Conference on Document Image Analysis for Libraries*, pages 133–144, 2006.

[14] F. Le Bourgeois and H. Emptoz. DEBORA: Digital accEss to BOoks of the RenAissance. *International Journal of Document Analysis and Recognition*, 9(2):193–221, 2007.

[15] X. Li, C. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *ACM International Conference on Multimedia Information Retrieval*, pages 180–187, 2008.

[16] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[17] G. Meng, N. Zheng, Y. Song, and Y. Zhang. Document Images Retrieval Based on Multiple Features Combination. In *International Conference on Document Analysis and Recognition*, volume 1, pages 143–147, 2007.

[18] G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, 2000.

[19] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte. Document Image Segmentation Using a 2D Conditional Random Field Model. In *International Conference on Document Analysis and Recognition*, volume 1, pages 407–411, 2007.

[20] J. Ogier and K. Tombre. Madonne: Document Image Analysis Techniques for Cultural Heritage Documents. In *EVA Conference on Digital Cultural Heritage*, pages 107–114, 2006.

[21] G. Park, Y. Baek, and H.-K. Lee. Re-ranking algorithm using post-retrieval clustering for content-based image retrieval. *Information Processing & Management*, 41(2):177–194, 2005.

[22] J. Ramel, S. Busson, and M. Demonet. AGORA: the interactive document image analysis tool of the BVH project. In *International Conference on Document Image Analysis for Libraries*, pages 145–155, 2006.

[23] Y. Rui, T. S. Huang, and S. Mehrotra. Content-Based Image Retrieval With Relevance Feedback In MARS. In *IEEE International Conference on Image Processing*, pages 815–818, 1997.

[24] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–18, 2006.

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

[26] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-Time Bag of Words, Approximately. In *ACM International Conference on Image and Video Retrieval*, 2009.

[27] M. Wang, K. Yang, X. Hua, and H. Zhang. Visual tag dictionary: interpreting tags with visual words. In *1st Workshop on Web-scale Multimedia Corpus*, pages 1–8, 2009.