Relevance feedback strategies for artistic image collections tagging

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

ABSTRACT

This paper provides an analysis on relevance feedback techniques in a multimedia system designed for the interactive exploration and annotation of artistic collections, in particular illuminated manuscripts. The relevance feedback is presented not only as a very effective technique to improve the performance of the system, but also as a clever way to increase the user experience, mixing the interactive surfing through the artistic content with the possibility to gather valuable information from the user, and consequently improving his retrieval satisfaction. We compare a modification of the Mean-Shift Feature Space Warping algorithm, as representative of the standard RF procedures, and a learning-based technique based on transduction, considered in order to overcome some limitation of the previous technique. Experiments are reported regarding the adopted visual features based on covariance matrices.

Categories and Subject Descriptors

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

Keywords

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

1. INTRODUCTION

Easy-to-use multimedia retrieval systems are a very important addition to museums, as well as very precious tool for researchers and experts. They can bring a phenomenal leap forward in the user experience on artistic collection, and the amount of artistic masterpieces available in various forms (paintings, books, manuscripts, even photos of sculptures and architectures eventually) represent a huge - but often unexploited - potential, both for users and public or private companies in the field of cultural heritage preservation and valorisation. The digitalization and all the modern technologies, like multitouch, can enclose the artistic work into a unique "paradigm", which allows freedom of data elaboration, security and replication features, at the same time gathering a great appeal to the public. Moreover, 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.

In this paper, we want to expand our previous work by focusing on the relevance feedback aspect. In fact, until now, the usual approaches on domain specific artistic collections (or digital libraries in general) 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, but a structured knowledge representation could result too complex, and the effort to design a specific ontology for each type of illuminated manuscript as well as the cost of manual annotation, could result not to be affordable for the museum owning the masterpiece. Thus our proposal is the adoption of tagging, which can exploit the existing commentaries and the interaction of users as a valuable alternative. The process of annotation by tagging, leveraged by the visual similarity search, can be further —and, based on our tests, strongly— improved using relevance feedback. So, the participation of the user in the retrieval loop not only is fundamental to dramatically improve the retrieval performance of the system, but also allows the user himself to boost his experience in the application, therefore his pleasure on using it.

After an overview of related work in Section 2, in Section 3 we will briefly describe the reference multimedia retrieval system (its main features and innovations), providing some hints about the motivation of this work and how the user interface changes accordingly. In Section 4 we will describe the main feature we used in our test, while the main topic of the work, that is the relevance feedback, is deepened in Section 5. Experimental results in Section 7 will precede some conclusion remarks and future works.

2. RELATED WORK

The use of relevance feedback strategies in information retrieval and in particular in content-based image retrieval

ICMR '11, April 17-20, Trento, Italy

systems is widely considered a very precious (sometimes necessary) to the system itself. Actually it is the most effective way to capture user's information need and, more generally, user's search intention. The reason is pretty straightforward: 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 the input of feedbacks (positive or negative). The literature on this topic is countless [23][4], since this problem can be faced from several point of view (computer vision, database management, human-computer interaction, artificial intelligence, even psychology). Moreover, aside the research on the algorithm for relevance feedback, there is a wide literature about the way in which the performance of a system with relevance feedback can be safely evaluated in order to provide fair comparison with different techniques. Regarding the algorithms, we can identify very generally three classes. In the first one (called Query Point Movement, QPM in short), we try to move the query point in order to create a more complete query (a fast technique to overcome to slow convergence is proposed by [11]). In the second one (called Feature Space Warping, FSW in short), we try instead to manipulating the feature space or the metric space, in order to shape it in the direction of the users' feedbacks [1][13]. The third one applies some machine learning procedures to learn how to separate relevant samples from irrelevant ones [17][18]. Among the usual techniques, based on SVM or boosting, we preferred testing a transduction-based learning. Some author followed the same path (see [21][16][15][14]). The idea is to take advantage both of the unlabeled and labeled samples in a transductive inference manner, learning from an incremental amount of training samples (feedbacks, in this case). Given the algorithm, the problem of evaluation is controversy. Even back in the Seventies, Williamson [20] proposed an evaluation methodology to tackle the so called "ranking effect" in the "fluid" relevance feedback evaluation, i.e. the overestimated performance improvement (in terms of recall and precision) due to the reposition of positive feedback in the top of the rank, aside the underestimated performance improvement of the "frozen" relevance feedback evaluation, which maintain the original ranks of documents along the sessions. In his "reranked original" ranking proposal, the best ranks are assigned to thy relevant documents and the worst ranks to the non-relevant documents; those documents not yet judged would remain in their original order, but with a rank decreased by the number of non-relevant documents identified. In [10] a comprehensive analysis tries to find out the reasons of relevance feedback evaluation problems, in particular problems with the dataset (characteristics and relative ground truth), problem with the comparison (different measures, different ranking approaches that make a comparison unfair, the need of rank normalization), and finally the problem of the parameter settings which can be impractical in real context. In our opinion, a quite fair and complete set of measures has been proposed in [12], where authors proposed:

- *actual* recall and precision (computed at each iteration and relative to the current set of retrieved images solely)
- *new* recall and precision (computed at each iteration and relative to the previous set of retrieved images solely)
- *cumulative* recall and precision (computed at each iteration and relative to the whole set iterations so far)

In this way, we can describe the behavior of the retrieval system both in terms of speed (how fast valuable images are retrieved over time) and in terms of completeness (how many good images the retrieval system finds out globally). Finally, as suggested in [9], we tried to concentrate the analysis con *feasible search task*, i.e. visual topics with a good number of representatives with a low degree of uncertainty in the evaluation, in order to assure a valuable reference ground truth. As mentioned before, the learning strategy for relevance feedback is very popular in literature.

3. THE VISUALLY ASSISTED TAGGING SYSTEM

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 [6]. 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 [5], 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 (a user going to the museum to explore digital manuscripts) to a very complex one in the last case (an expert who want to study the digital manuscript), in order to satisfy the different degrees of expressive power needed. In this paper, we want to focus more on the last use case, and in particular on the activity of a "art curator" which wants to provide the system the amount of textual information (like tags or commentaries) to form the basic corpus of expert value available for the user. Since we believe that basically normal users and experts are just users, conceptually acting in very similar ways apart from the level of complexity of their queries, we are proposing to use the same techniques also

to facilitate the annotation work of the curator, by leveraging and improving the relevance feedback (and partially the user interface proposed in [2]). We believe that the relevance feedback, being the direct entry point of the user's needs in the system, is the key component to be improved in this context.

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 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.

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 needs. For this reason, if the proposed user interaction is good and effective for his purpose, the user will be pleased to come back using the application. Starting from the proposal in [2], we introduced a new user interface specifically designed to facilitate the work of a curator, or generally the work of a user interested on a complete annotation of the artistic work. Some screenshots are provided in Fig. 1.

In 1a, the list of $k = 25$ most similar images are presented to the user. The selection of multiple pictures represent the "positive" feedbacks. Since statistically the number of "negative" pictures is consistent (especially when the recall is close to 1), we considered as "negative" feedbacks all the pictures left unselected. As we will detail better in Section 5, this choice could cause the relevance feedback technique we used in [2] to be less effective: for this reason, in this work we explored a couple of interesting alternatives to solve the problem. When the search is concluded, the tagging interface 1b allows an easy and fancy association of tags shared between all retrieved images.

4. 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 $\mathbf{Z}$ be the feature point matrix where $\{\mathbf{z}_i\}_{i=1..N}$ is 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

$$\mathbf{C}_I = \frac{1}{N-1}(\mathbf{Z} - \mu)(\mathbf{Z} - \mu)^{\mathbf{T}} \tag{2}$$

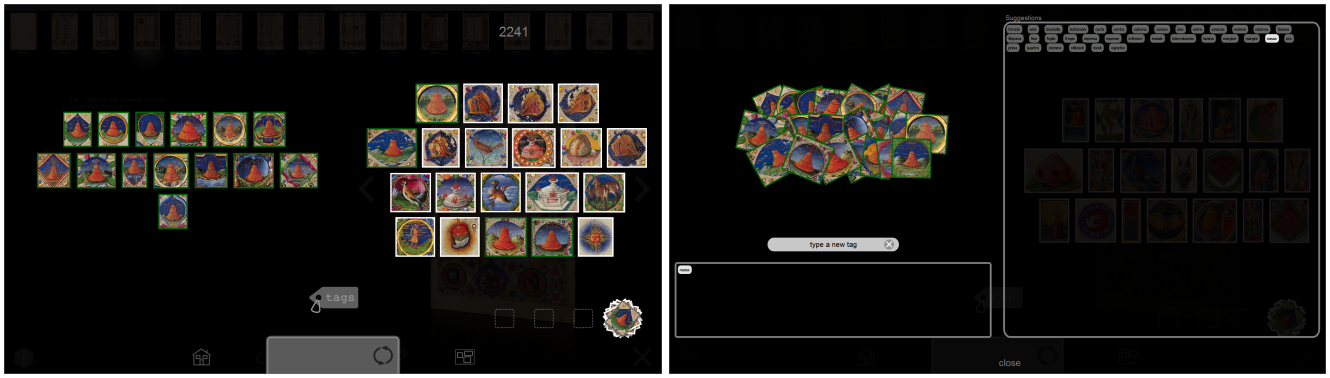
where μ is the column vector of the mean of the feature points. The noise corrupting individual samples is largely filtered out with the average filter during covariance computation. The descriptors are low-dimensional, and due to symmetry, C_I has only $(d^2 + d)/2$ different values.

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 $\phi(I, x, y)$

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

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 [7] the following distance measure for positive definite symmetric matrices is



(a) (b)

Figure 1: The new UI introduced to easy up the work of the curator in the process of complete tagging of the visual content.

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[19]: 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 the selection of an appropriate projection origin, every covariance matrix gets projected to a 28-dimensional feature vector laying on an Euclidean space.

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

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

5. RELEVANCE FEEDBACK

The use of relevance feedback in multimedia retrieval system has a double advantage. In the first place, a significant corpus of literature works proved that relevance feedback is a powerful solution to improve the performance of the retrieval system; in the second place, the participation of the user in the retrieval loop allows to improve his experience within the application, therefore his pleasure on using it.

5.1 Mean-Shift Feature Space Warping

In this work, we started from the relevance feedback technique proposed by Chang *et al.* [3] called Mean Shift Feature Space Warping. 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

$$\mathbf{p}' = \mathbf{p} + \lambda \sum_{j=1}^{M} u_j \exp\left(-c|\mathbf{p} - \mathbf{f}_j|\right)(\mathbf{w} - \mathbf{p}) \tag{10}$$

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 $\mathbf{w}$ 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

Figure 2: Example of pictures grouped by class. (a) is identified with "nassa" and represents ancient fish trap; (b) is identified with "Fido" and represents a symbol of the Estense family; (c) is identified with "rosa", and represents a rose inside a ring with a diamond, which is the symbol of Duke Ercole I of Este; (d) is identified with "bacinella" and represent a fountain with flames; finally (e) is identified with "stemma" and represent a symbol of the Estense family.

adopt the Rocchio's query movement formula:

$$\mathbf{w}' = \alpha\mathbf{w} + \beta\overline{\mathbf{f}_p} - \gamma\overline{\mathbf{f}_n} \tag{11}$$

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 (Eq. 8). 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 show them to the user in order to perform further refinements.

5.2 Mean-Shift Feature Space Warping with Remapping

Since this mapping is a homeomorphism around the neighborhood of the point, the structure of the manifold is locally preserved. The problem here is that the first step of MSFSW moves the warping center away from the current one, and this may impact on the quality of the projected vectors. For this reason we propose to employ an intermediate step of reprojection around the new warping center. The proposed relevance feedback approach works iteratively according with the following sequence of steps:

1. Given the previous warping center $\mathbf{w}_{i-1}$ and feedbacks $\{\mathbf{f}\}_{i-1}$, the new warping center $\mathbf{w}_i$ is computed by means of Eq. 11;

2. All points $\{\mathbf{p}\}_{i-1}$ and $\mathbf{w}_i$ are reprojected on the manifold, defining the set of remapped points $\{\mathbf{R}\}_i$, exploiting Eq. 9:

$$\mathbf{R}_i = \mathbf{W}_{i-1}^{\frac{1}{2}} \exp\left(\mathrm{vec}_{\mathbf{I}}^{-1}\left(\mathbf{p}_{i-1}\right)\right) \mathbf{W}_{i-1}^{\frac{1}{2}} \tag{12}$$

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

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

4. At this point, we can apply the FSW on the set $\{\mathbf{r}\}_i$, as in Eq. 10, 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 (for step 1). Thus in step 2, 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.

Moreover, we observed that for our purposes, the use of negative samples as well as the presence of the previous warping center in Eq. 11, can have a negative impact on the performance of the algorithm. This is especially true when the discriminative power of the used feature is low for a particular class of images, or when the user selects some images as negative feedbacks because these contain semantically different concepts, despite the visual appearance is 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 the mean of the positive feedbacks collected so far. In this way we can also simplify the system getting rid of three additional parameters (α,β and γ).

6. TRANSDUCTIVE RELEVANCE FEEDBACK

As mentioned in the Section 2, the relevance feedback problem can be analyzed as a semisupervised learning problem, in which the positive and the negative feedbacks given

by the users constitute iteratively (and incrementally) the training set of the algorithm. In this paper, we propose a graph-based transductive learning method to tackle this purpose, defining a graph where the vertices represent the labeled and unlabeled images of the dataset, while the edges incorporate the similarity between them, in our case obtained from the distance between covariance matrices. Graph-based methods are nonparametric, discriminative, and transductive by definition [22], and labels can be assumed to be smooth over the graph. Starting from the whole dataset with n images, let's define a set $\mathcal{L}$ of labeled images $(x_1, y_1), \ldots, (x_l, y_l)$ in which C classes are defined, so $y_i \in 1 \ldots, C$. The other images belongs to a set U of u unlabeled images, with $n = l + u$. Now let's define a function $f : \mathbb{R}^n \to [0, 1]$ which denotes the confidence of each image to one class. Formally, we can define a cost function J on f as:

$$J(f) = \sum_{(i,j)=1}^{n} \|f(x_i) - f(x_j)\|^2 w_{ij} + \lambda \sum_{i=1}^{l} \|f(x_i) - y_i\|^2 \tag{14}$$

with λ as a regularization parameter (in our case, $\lambda = 1$). This equation, in a minimization process, tries to match the confidence w_{ij} between samples with the the true confidence y_i with respect of the current confidence. Once converted in matrix notation, Eq. 14 becomes:

$$J(f) = (f(X) - Y)^T (f(X) - Y) + \lambda f(X)^T L f(X) \tag{15}$$

where $L = D - W$ is the graph Laplacian. W is the weight matrix, while D is the matrix which represents the degree of vertices:

$$d_{ii} = \sum_{j=1}^{n} w_{ij} \tag{16}$$

The cost function minimization has a closed form solution, that is:

$$f = (I - \lambda L)^{-1} Y \tag{17}$$

At the end of the computation, f contains the class confidence of each new sample.

In the transductive process, we want to transfer labels from labeled samples to unlabeled ones: in other words, we want the samples which are close in the feature space to share the same label. To satisfy this *local constraint*, we construct accordingly the weight matrix W, that is the matrix in which each element w_{ij} contains the relation between two vertices (thus two images) xi, xj. To move from distances to affinities, we use the following formulation:

$$w_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{18}$$

where σ is a bandwidth parameter to tune the relations between vertices, and the distance is computed as the L_2 norm after the conversion of covariance matrices in the Riemannian Manifold to vectors in the Euclidean Space. W can be subdivided into four submatrices:

$$W = \begin{pmatrix} W^{ll} & W^{lu} \\ W^{ul} & W^{uu} \end{pmatrix} \tag{19}$$

where W^{ll} (a full connected graph) denotes relations between labeled data, W^{uu} (a k-nearest neighbor graph) denotes relations between the candidate images yet to label, and the symmetric subgraphs W^{ul} and W^{lu} (still k-nearest neighbor) denote the relations between positive and candidate images. The relations used to compute values in these graphs are the following:

$$W^{ll}_{i,j} = \frac{1}{n} \tag{20}$$

$$W^{uu}_{i,j} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2_{uu}}\right) & \text{if } x_i \in knn(x_j); \\ 0 & \text{otherwise;} \end{cases} \tag{21}$$

$$W^{lu}_{i,j} = W^{ul}_{i,j} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2_{uu}}\right) & \text{if } x_i \in knn(x_j); \\ 0 & \text{otherwise;} \end{cases} \tag{22}$$

where $k = 10$ and $\sigma = 1$.

After the first ranking by similarity, the user selects the positive feedbacks while the unselected samples are considered as negative. Then the process described in this section is iterated following the user's need or until no more changes in the rank occurs.

7. EXPERIMENTAL RESULTS

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. 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 [8], followed by a manual correction of the extracted pictures, we obtained a dataset of 2282 pictures. We performed an automatic simulation of relevance feedback interaction, in order to avoid human errors. We considered 6 distinctive classes defining a total amount of 171 queries over the dataset (one for each training sample, to evaluate the performance of the same query starting from different prototypes). For each query, results can be safely considered objective from both a semantic and a visual point of view. We compared the following algorithms:

- *Naive relevance feedback* (actually no relevance feedback at all): the system discards the current set of n results and propose to the user the next n, following the original rank given by the visual similarity;
- *MSFSW*: the original Mean Shift Feature Space Warping proposal by [11], with an empirically optimized set of parameter $\alpha = 0.2$, $\beta = 0.5$ and $\gamma = 0.3$ for the means-shift part and $\lambda = 0.7$ and $c = 0.8$;
- *MSFSW with Remapping*: our modification of the original MSFSW algorithm which performs a remapping of the entire feature set using the mean of positive feedbacks as tangent point for the conversion from Riemannian Manifold to Euclidean Space;
- *TL*: the transductive learning approach which uses positive feedbacks as samples and define the relevance

feedback as a process to assign a label to unlabeled samples. The affinity matrix is filled only for the $k = 20$ nearest neighbors.

The performance has been evaluated in a user-centric perspective: we chose to use metrics clearly comprehensible by a user in front of the application, and we included a fixed number $T = 10$ of iterations, to convey that the user will get bored and stop pursuing in the search after 10 refinements. We indicate as *step* the average number of steps to get the maximum retrieved images (within 10 iterations), and we indicate as $@i$, with $i = 0 \dots T$ the incremental recall, namely the recall provided by the system at each step i. While the first steps gives an idea of the convergence capabilities of the algorithm, the last one at the last step gives an overall metric of the algorithm itself. The results are presented in Table 1.

Using the naive technique as baseline, which shows poor performance (thus an increasing amount of work for the curator in order to tag the entire collection), we can see that the original MSFSW technique provides satisfying results in a casual user perspective: in fact, during the initial iterations, a good amount of positive feedbacks is pulled up in the first positions of the rank. In the original interface proposed in [2], by dragging pictures in green and red pots, the user can select in an emotional way which sample to consider positive and which to consider negative. But if we want to provide a mass tagging functionality, thus focusing only on positive feedbacks and assuming that the others are all negative, aside the inefficient user interface design, the original algorithm suffers, because the prevailing amount of negative samples induce the query point movement part and later the feature space warping to push away from the query a lot of good pictures yet to retrieve. The modification proposed in this paper, which neglects but does not weight the negative samples in the query point movement, allows the algorithm to focus solely on the good samples retrieved, and to arrange the feature space in such a way to favor the rise up of good pictures in the ranking. This behavior is consistent along all the queries. The transductive learning shows overall a lower performance in terms of convergence (it tooks some more iterations to reach the same level of recall of MSFSW) but it shows also the capability to reach an higher recall at last. This behavior is consistent with the common benefits and drawbacks connected to the use of a supervised learning procedure: the algorithm is capable of describing more efficiently the positive pattern, thus conveying a better recall at each refined classification step, but when the initial ranking by similarity is not able to provide a sufficient number of positive samples to learn from, due to a reduced effectiveness of the visual feature for that particular class or that particular prototype of the class, the algorithm has not sufficient information to move on the learning process, and it behave just the same as the naive feedback, solely moving towards the top the following samples of the rank. Moreover, this approach - at least in its original formulation - is quite slow and not very scalable with the increase of the size of the dataset (due to the matrix computations). Some solutions (approximations and optimizations) have been already proposed, and we will tackle them in the future.

Table 1: Comparison of the 5 relevance feedback proposals. For each class *cls*, the step @0 represent the recall after the first query-by-sample ranking, using the visual feature only. The following steps gives an idea of the improvements provided by the relevance feedback techniques at incremental steps. The last column *step* reports the mean value of the step in which the maximum recall (within T iterations) is reached

cls	@0	@1	@2	@3	@4	@5	@10	step
0	17.82	30.9	39.7	46.5	50.5	53.5	62.5	9.24
1	8.18	13.9	17.1	21.0	23.8	25.9	36.4	8.36
2	14.06	21.8	27.4	32.2	35.6	39.5	48.5	8.24
3	20.27	29.9	35.1	38.8	41.0	42.3	49.1	8.54
4	14.13	19.9	25.5	28.8	31.3	35.7	43.2	7.32
5	13.19	22.5	27.2	31.3	35.0	37.4	44.4	9.08
all	14.34	23.1	28.6	33.1	36.3	39.0	47.4	8.58

(a) Naive Relevance Feedback

cls	@0	@1	@2	@3	@4	@5	@10	step
0	26.82	49.7	57.6	62.6	69.5	74.0	87.5	8.03
1	10.26	18.1	20.4	21.5	22.5	23.4	44.8	9.50
2	19.95	31.3	36.3	36.7	36.7	37.0	38.3	4.48
3	26.33	42.0	48.5	49.1	49.1	49.3	49.3	2.88
4	19.67	31.9	36.3	36.6	36.6	37.7	40.7	6.00
5	19.91	34.5	42.2	43.6	44.8	46.8	51.5	7.78
all	20.20	34.5	40.3	41.9	43.7	45.4	54.1	6.85

(b) Mean-Shift Feature Space Warping

cls	@0	@1	@2	@3	@4	@5	@10	step
0	26.82	55.5	77.0	89.8	92.9	94.0	96.7	6.47
1	10.26	23.7	36.1	45.4	50.3	52.9	61.8	8.11
2	19.95	32.2	42.9	45.4	45.8	46.0	46.3	3.71
3	26.33	38.8	50.9	62.1	68.5	70.9	71.2	4.35
4	19.67	35.7	43.2	49.3	51.2	53.7	58.4	5.79
5	19.91	33.3	49.2	58.7	62.1	63.8	65.4	6.11
all	20.20	36.6	50.8	59.9	63.5	65.3	68.6	6.01

(c) Mean-Shift Feature Space Warping with Remapping

cls	@0	@1	@2	@3	@4	@5	@10	step
0	26.82	42.4	56.5	68.8	76.1	83.5	91.3	7.44
1	10.26	16.0	20.8	24.9	27.6	29.6	39.1	8.75
2	19.95	29.9	37.2	45.4	52.8	56.2	59.0	5.57
3	26.33	37.4	46.4	55.6	65.5	74.4	98.5	9.96
4	19.67	26.9	31.3	32.7	33.8	36.3	38.2	4.26
5	19.91	31.1	41.9	49.9	56.7	62.1	73.8	10.31
all	20.20	30.5	39.3	46.8	52.8	57.8	68.0	8.12

(d) Transductive Learning

8. CONCLUSIONS

In this paper we improved our design process and research about an innovative and complete system for the exploration and annotation of artistic image collections. In particular, we focused on the use of relevance feedback in a user-centric perspective, in order to propose a solution for the problem of the mass tagging of an entire image collection database (in this case, pictures extracted from an illuminated manuscript). The procedure described shows the advantages connected to the use of an effective relevance feedback technique, in terms of amount of time and cost for a curator to annotate comprehensively the entire dataset.

9. 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.

10. 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] D. Borghesani, C. Grana, and R. Cucchiara. Surfing on artistic documents with visually assisted tagging. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1343–1352, New York, NY, USA, 2010. ACM.

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

[5] 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.

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

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

[8] C. Grana, D. Borghesani, and R. Cucchiara. Automatic segmentation of digitalized historical manuscripts. *Multimedia Tools and Applications*, pages 1–24, July 2010.

[9] M. J. Huiskes and M. S. Lew. Performance evaluation of relevance feedback methods. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 239–248, 2008.

[10] X. Jin, J. French, and J. Michel. Toward consistent evaluation of relevance feedback approaches in multimedia retrieval. In *International Workshop on Adaptive Multimedia Retrival*, July 2005.

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

[12] J. Luo and M. A. Nascimento. Content-based sub-image retrieval using relevance feedback. In *ACM International workshop on Multimedia databases*, pages 2–9, 2004.

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

[14] V. Radosavljevic, N. Kojic, G. Zajic, and B. Reljin. The use of unlabeled data in image retrieval with relevance feedback. In *Symposium on Neural Network Applications in Electrical Engineering*, pages 21–26, Sept. 2008.

[15] H. Sahbi, J.-Y. Audibert, and R. Keriven. Graph-cut transducers for relevance feedback in content based image retrieval. In *IEEE International Conference on Computer Vision*, pages 1–8, Oct. 2007.

[16] H. Sahbi, P. Etyngier, J.-Y. Audibert, and R. Keriven. Manifold learning using robust graph laplacian for interactive image search. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

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

[18] K. Tieu and P. Viola. Boosting image retrieval. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 228–235, 2000.

[19] 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.

[20] R. E. Williamson. Does relevance feedback improve document retrieval performance? *SIGIR Forum*, 13:151–170, May 1978.

[21] Y. Wu, Q. Tian, and T. Huang. Integrating unlabeled images for image retrieval based on relevance feedback. In *International Conference on Pattern Recognition*, volume 1, pages 21–24, 2000.

[22] Y. Zha, Y. Yang, and D. Bi. Graph-based transductive learning for robust visual tracking. *Pattern Recognition*, 43:187–196, January 2010.

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