RELEVANCE FEEDBACK AS AN INTERACTIVE NAVIGATION TOOL

Daniele Borghesani, Costantino Grana and Rita Cucchiara
Università degli Studi di Modena e Reggio Emilia
{*name.surname*}*@unimore.it*

Keywords: CBIR, HCI, artistic collections, relevance feedback

Abstract: Image collections are searched in common retrieval systems in many different ways, but the typical presentation is by means of a grid styled view. In this paper we try to suggest a novel use of relevance feedback as a tool to warp the view and allow the user to spatially navigate the image collection, and at the same time focus on his retrieval aim. This is obtained by the use of a distance based space warping on the 2D projection of the distance matrix.

1 INTRODUCTION

The growing availability of multimedia content, especially pictures, and the proposal of increasingly efficient content analysis techniques led to the development of impressive multimedia systems in recent years. Nevertheless there is a significant gap between the research view of such systems and the user perspective, which is strongly influenced by the way information is presented. This led to a standardization of the interface solutions, based on their success On the market.

Let's look at the way in which image library tools usually present information to the user. For example, in desktop applications like iPhoto or Picasa, images are classified using default metadata like GPS, tags possibly associated to pictures, and time stamps. With this simple information, the system can perform an automatic grouping of data to assist the user in the process of management of his library. Another quite standard functionality is the filtering, using both metadata or —very rarely— rough visual information based on color. All these functionalities finally relies on a very standard grid-based layout representation, which is very familiar to most of users but, especially in a similarity retrieval context, can be considered a bad design choice since it erases all the similarity relations (connections) between images. The same kind of problem is clearly recognizable in all the most used web search engines, like Google, Bing or Yahoo. Only very recently Google introduced visual search capabilities (subject to features precomputation), which seems quite good on specific objects retrieval, but behaves much more inaccurately on average. In the majority of situation, instead, we cannot search using an image as a query, but we need to start from a standard textual input. Secondly, when we have the resulting list of pictures, a grid-based layout is proposed to the user. In every case the modus operandi is just the same: look and scroll for more images. We believe that this approach is essentially flawed, because of two main reasons: it does not convey visual feedback about the content of the collection, and it does not dynamically react to the feedbacks of the user.

In this paper, we want to propose an easy solution to solve this interface gap. Starting from a solid set of content analysis and indexing techniques (which can be eventually designed to fit the large scale requirements), we propose the relevance feedback not only as an effective tool to improve the raw performance of the retrieval system, but mainly as a mean to help the user navigating into the collection, especially when no metadata are available or when the the search intentions cannot be easily expressed as textual queries. In this way, we want to facilitate the user in the process of manipulation of the information: by visually surfing through images, the user can build connections and feel emotionally involved in the navigation experience, using the relevance feedback to warp the space around his needs, quickly learning the results content and possibly moving to a destination he did not even think about when he started. We believe that, in the near future, the similarity search will have a key role in the market, not in order to substitute the search by text but more importantly in order to complement

it. Actually it will probably be next door feature of image library management tools and web search engines, complementing other research efforts focusing on classification, annotation and so on.

2 BACKGROUND

The problem of image retrieval is two-fold. In the first place, we need fast and effective techniques to convey visual similarity to the user. In the second place, we need an effective technique to allow the user to manage the results.

Regarding the first problem, a great amount of literature has been proposed. Among it, we think that the natural choice is a global feature representation, providing a compact summary by aggregating some information extracted at every pixel location of the image. The bag-of-words approach, a global representation build of clustered local features like SIFT (Lowe, 2004) or SURF (Bay et al., 2008) as a visual dictionary, is generally considered the state of the art. For a complete comparison of performance of local features in CBIR, please refer to (Mikolajczyk and Schmid, 2005). Most of these local descriptors use luminance information only. Nevertheless, both color and shape are widely considered important visual characteristics in a cognitive context, so an interesting way to account this information is by using the *covariance region descriptor*, proposed by Tuzel *et al.* in (Tuzel et al., 2008), which aggregates the correlations of a custom amount of elementary sources of information (like color, shape, spatial information, gradients). Moreover, great interest was devoted to GIST feature, a statistical summary of the spatial layout properties (Spatial Envelope representation) of the scene (Oliva and Torralba, 2006).

To solve the second problem, as pioneered by Renninson in (Rennison, 1994), a presentation strategy is required. The classical spatial arrangement of images is their placement on a grid, typically in row-major ordering based on relevance. Despite its simplicity, this visualization is unable to convey information on the structure of the collection, for example the availability of a cluster of similar images. As described in (Heesch, 2008), alongside with more standard approaches based on static hierarchies or clustering, the main approaches are build around a network based or a dimensionality reduction based representations. Multi-Dimensional Scaling (MDS) solves a non linear optimization problem by determining the mapping that best approximates the high-dimensional pairwise distances between data points. One of the initial proposals was the Sammon mapping by (Sammon, 1969). An interesting proposal of this kind is the Hyperbolic-MDS by (Walter, 2004), which exploits the hyperbolic space $\mathbb{H}^2$ to map the most significant images in the center of the projection (thus visualizing them with a greater detail) while displacing the others along the curve $\mathbb{H}^2$ falling towards infinity with a smaller scale; moreover this projection has the advantage of allowing to focus the view in different points by applying the Möbius transformation. A number of other non-linear projections have been proposed to solve the prohibitive computational costs, for example the isometric mapping (ISOMAP) (Tenenbaum et al., 2000), the stochastic neighbor embedding (SNE) (Hinton and Roweis, 2002) and the local linear embedding (LLE) (Roweis and Lawrence, 2000). An older yet effective approach, especially in large scale contexts, is finally the FastMap (Faloutsos and Lin, 1995) which exploits a set of pivot objects to project points in the reduced space. This technique, exploited also in this paper, has the advantage to allow easily a fast insertion of new objects within the map.

3 RELEVANCE FEEDBACK FOR IMAGE SURFING

The first task in image searching on large scale collections is clearly managing the scalability problem. Many techniques for approximated nearest neighbor (ANN) search, starting from the LSH (Andoni and Indyk, 2006) up to the product quantization (Jégou et al., 2011), allow to greatly improve the performance using vocabulary codes (with precomputed distances) in place of real features. Moreover image search based on contextual information (as done by all search engines) proves to be definitely effective. The real limitation of todays multimedia systems is within the interaction possibilities.

The most important way in which the user can help the system cross the semantic gap and interact with the retrieval results, i.e. the relevance feedback, becomes first of all prohibitive in large scale contexts. Just consider the usual approaches: query point movement (QPM), feature space warping (FSW) or machine learning approaches (Chang et al., 2009). QPM notoriously suffers of slow convergence, and does not guarantee to find intended targets; a fast QPM technique, trying to fix this problem, has been proposed by (Liu et al., 2009). FSW requires a full space re-encoding, and no proposals at the best of our knowledge take into account FSW in large scale scenarios. Finally the learning is notoriously a heavy procedure, often requiring an offline processing and hardly capa-

ble of producing real time results. Moreover, the relevance feedback is proposed to the user as a tedious procedure (as well as the annotation) to overcome the limitations of the system itself, which could be considered an admission of poor quality.

Nevertheless, the ability to guide the system towards the desired result needs to be considered as an important feature. The user himself implicitly demands this kind of capability, because visual similarity is mostly helpful when the user does not clearly know or is not capable of expressing the subject of his search: as a matter of facts if he could, he would type the precise query on the search engine. This is even more true when the user is approaching the image collection for fun or curiosity: in this scenario the user is mainly interested in surfing through pictures being guided by his emotional preferences, using visual cues as exploration rails. In the meantime, new and refined results could be suggested by the retrieval system, adjusting his search goal.

In order to satisfy all these requirements, we need to visualize the effect of relevance feedbacks from the original feature space into the two-dimensional mapping. This procedure allows the system to show to the user a real-time feedback of his manipulations, bringing him into the collection itself.

We need to provide the user with a first 2D visualization of his query results. The technique used in this step is FastMap, due to its high performance and the ability to quickly include new points to the map without recomputing the entire mapping. This algorithm briefly works as follows (Faloutsos and Lin, 1995). Firstly, two distant-enough objects are chosen with an heuristic approach. Given a distance function $\mathcal{D}()$ between each pair of objects O_a and O_b in the feature space, each object O_i is projected to object $O_i^{'}$ on the line joining the pivots (O_a, O_b) using the cosine law and obtaining the x coordinates. Then the y coordinate is computed using the distances $\mathcal{D}'$ on the hyperplane perpendicular to the line (O_a, O_b). These may be obtained from the original distance $\mathcal{D}$ by means of Eq.1:

$$\mathcal{D}'\left(O_i^{'}, O_j^{'}\right)^2 = \mathcal{D}(O_i, O_j)^2 - (x_i - x_j)^2 \qquad (1)$$

When the process is completed, the pictures are visualized on the two-dimensional plane adjusting the scale.

When a query O_q is selected by the user, the points are adjusted in order to support the similarity ranking. In particular the user requires a new projection which better reflects the distances from the query, thus the angle of points from the query is kept fixed, while the distance is scaled along the unit vector proportionally to the ranking itself. In this way, the similar pictures get closer to the query, while the dissimilar ones are moved away. At this point, the user is focused on the query itself (at the center of the screen) and the most similar content within the results is placed nearby, easily gathering his attention.

The user can now provide feedbacks on the results, highlighting what he likes (being more similar to the query he submitted) and what he dislikes (being different from what he expects). For each point O_i in the results set, the system finds the nearest element of both positive and negative feedbacks sets (a process which can be eased up with approximate search) and warps the space. In particular, given f_p the distance from its nearest good feedback (including the query image) and f_n the distance from its nearest bad feedback, the system computes the distance for the projection $\mathcal{P}$ as:

$$\mathcal{P}(O_i, O_q) = \mathcal{D}(O_i, O_q)\left(1 + \frac{f_p - f_n}{\max(f_p, f_n)}\right) \qquad (2)$$

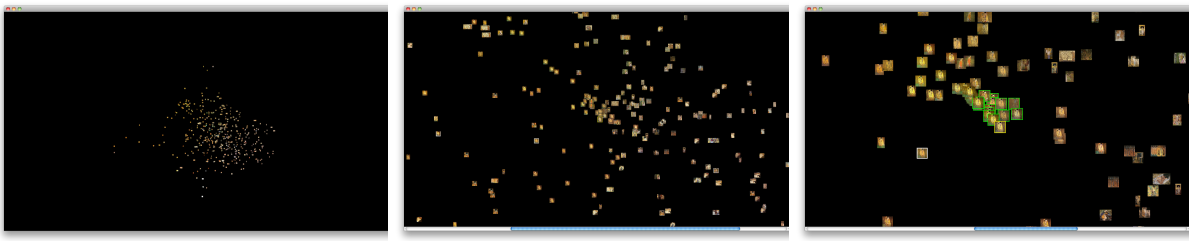
The equation states that what is positive should be moved towards the query, while what is negative should be pushed away. The "positiveness" of an image is related to how much more similar to a positive than to a negative the image is. The images may now be ranked according the warped distances and the visualization is updated by moving the images along the line which connects the points to the query in the 2D plane. The new distances are ordered according to the ranking.

Compared with other relevance feedback approaches, this solution may perform worse with respect to the global recall or precision. The real merit, which becomes essential, regards the interface aspect: in fact the changes induced to the ranking are limited to the local neighborhood of the selected feedback element. In other words, only the points for which the feedback is the nearest positive or negative feedback are influenced, therefore a strong connection between the visual mapping and the observed changes appears. Moreover the use of a ranking based projection has the effect of showing the similar images slowly approaching the query, thus the user's attention focus.

The user is still allowed to move the images as he feels like, implicitly asking to prevent the image from being moved by the automatic positioning. Note that the distance calculations are always performed on the original distances, so removing a feedback allows to step back to the previous position: this is an easy way to "undo" the user's choices.

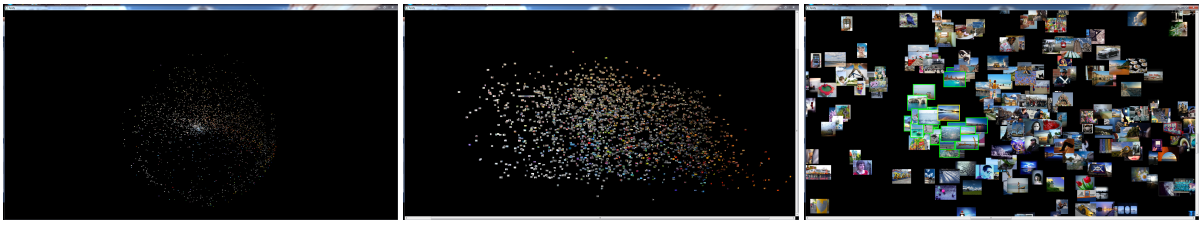
4 SAMPLE APPLICATIONS

One of the most immediate implementations of this approach is the image web search interface. Cur-



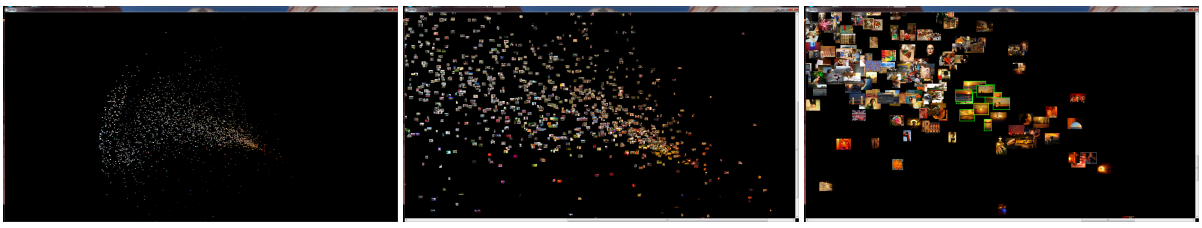
(a) (b) (c)

Figure 1: Application example with the Google query “klimt”.



(a) (b) (c)

Figure 2: Application example by query one of the ImageCLEF images, specifically representing sea pictures.



(a) (b) (c)

Figure 3: Application example by query one of the ImageCLEF images, specifically representing sunset pictures.

rently, Ajax based interfaces are common, so the presentation could be further enhanced to support our idea. The web server could provide image features or they could be directly computed by the client interface if they are simple enough. Suppose we use Google Images to look at something related to a painter ("klimt" could be our query term): the set of images is presented in a 2D mapping (Fig. 1(a)) and the user quickly pans and zooms through the results, not dissimilarly from what he would do scrolling down the results page. After identifying some interesting content, he can select it (Fig. 1(b)) and then better convey his interest with further refinement, which are likely to attract other versions of the painting and similar ones, while repelling unrelated results (Fig. 1(c)). Note that this is definitely an ephemeral interest, exactly related to the moment and the feelings of the person: it is likely that a new object of interest gets identified by the collection exploration, to start over again.

The same approach can be extended quite easily to surf through wide collections of images. The idea to allow the user to zoom into the dataset, and filter out the pictures based perceptual similarity exploiting interactive relevance feedback, can be a winning key in the process of managing efficiently large amount of data, focusing on the user's search intentions. Let's look, for example, the dataset provided for the CLEF Photo Annotation task (Nowak et al., 2011), aimed at the automatic annotation of a large number of consumer photos with multiple annotations. The user is presented with the 2D mapping of the images, computed in real-time by FastMap, and allowed to zoom and navigate through it(Fig. 2(a),3(a)). After identifying an interesting image (a sea landscape in Fig. 2, a sunset in Fig. 3), the user selects it and the other images are rearranged to convey their distance in the feature space from the selected query (Fig. 2(b),3(b)). This shows how well the 2D mapping is able to respect the original distance matrix. Now the user may simply select positive or negative samples, getting an immediate feedback of the effect of his choice on the mapping: selecting a negative feedback forces the image and some other neighbors to be pushed away and at the same time all the lower ranked image to be dragged toward the query. The selection of a positive feedback "recalls" images from outside the current view towards the query. Once the filtering is completed, the resulting images constitute a bucket of interesting images to be used somehow or annotated with a tag (Fig. 2(c),3(c)).

5 INTERACTION DESIGN REMARKS

Usually, the user interface design is considered the last step to deal with in the development of a retrieval application or a system, because the engine is considered the only real focus of the problem. Instead, we argue that in the next future of user centric application, the interface will become the key point of the system.

A compelling user interface, containing useful and engaging interaction paradigms, 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 example, Jaimes and Sebe (Jaimes and Sebe, 2007) show how to deal with user's emotional expressions as part of the data processing. These concepts evolved through time becoming what generally could be defined as natural interaction (Baraldi et al., 2009), exploiting means which are considered natural since they belong to the nature of human beings themselves.

The simpler and the more natural (let's say intuitive) the machine interaction is, the less amount of cognitive effort is delegated to humans. Nevertheless the design problem is remarkable. If we focus our attention on functionalities to be provided to users to accomplish some tasks, we risk losing the focus on intuitiveness. On the other side, an extreme simplification can lead to poorly performing functionalities. For this reason, we need to design these two aspects at the same time, linking very closely the search engines and the visualization techniques with the functionalities. If we design an effortless interaction capable of expressing naturally all the technically complex tasks to search, visualize and browse, we are close to a natural multimedia system really centered on user's desires and therefore really useful to him.

The aim of natural interaction is therefore the design of an interaction system able to getting rid of computer-friendly interaction paradigms (like windows, menus, scrollbars, mouses) towards more human-friendly paradigms. In this context, very important roles are played by concepts like aesthetic beauty, emotions and a playful dimension between the user and the system; moreover, an intensive use of animations and dynamic mathematical models is necessary in order to link the virtual interface with real life metaphors. Finally, the spatial organization of information is fundamental to improve content understanding, for example by clustering similar objects.

This proposal just moves towards this kind of interaction. The image collection is not only a list of images, but becomes a space to explore, reacting dy-

namically on the user's preferences collected continuously through relevance feedback. The entire system can be easily improved with convenient multitouch gestures. The removal of one or more undesired pictures can be triggered with swipe gestures, while the pinch gesture can allow to zoom the collection to focus on the individual pictures (or groups of pictures). Groups of good or bad feedbacks can be selected drawing circles around them. Once the collection has been filtered, according to the desired predominant visual characteristic, a tag could be associated to the resulting group of pictures, performing a visually assisted tagging.

6 CONCLUSIONS

In this paper we introduced a novel proposal for the presentation of image collections, obtained by querying or similarity search. We believe that the combined use of 2D mapping and relevance feedback allows the user to better express his querying intention, therefore easily surf through the results.

This technique, however much simple, could open a wide range of improvements of today's web search engines and image collections management software. For example, new results could be dynamically added to the mapping, based on the already selected images, thus formulating a new query based on the positive and the negative selections. Moreover, the visual similarity search can be exploited also to mine the not indexed content using positive feedbacks as suggested prototypes for the retrieval system. Finally, an interesting possibility is the exploitation of such an interactive experience to collect user provided information and therefore improving the retrieval system itself.

REFERENCES

Andoni, A. and Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *IEEE Symposium on Foundations of Computer Science*, pages 459–468.

Baraldi, S., Bimbo, A. D., Landucci, L., and Torpei, N. (2009). Natural interaction. In *Encyclopedia of Database Systems*, pages 1880–1885.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Comput Vis Image Und*, 110(3):346–359.

Chang, Y., Kamataki, K., and Chen, T. (2009). Mean shift feature space warping for relevance feedback. In *IEEE Image Proc*, pages 1849–1852.

Faloutsos, C. and Lin, K.-I. (1995). Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *ACM SIGMOD International Conference on Management of Data*, pages 163–174.

Heesch, D. (2008). A survey of browsing models for content based image retrieval. *Multimed Tools Appl*, 40:261–284.

Hinton, G. E. and Roweis, S. T. (2002). Stochastic neighbor embedding. In *Neu Inf Pro Syst*, pages 833–840.

Jaimes, A. and Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Comput Vis Image Und*, 108(1-2):116–134.

Jégou, H., Douze, M., and Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE T Pattern Anal*, 33(1):117–128.

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

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *Int J Comput Vision*, 60(2):91–110.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE T Pattern Anal*, 27(10):1615–1630.

Nowak, S., Nagel, K., and Liebetrau, J. (2011). The clef 2011 photo annotation and concept-based retrieval tasks. In Petras, V., Forner, P., and Clough, P. D., editors, *CLEF (Notebook Papers/Labs/Workshop)*.

Oliva, A. and Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Visual Perception, Progress in Brain Research*, 155.

Rennison, E. (1994). Galaxy of news: an approach to visualizing and understanding expansive news landscapes. In *ACM symposium on User interface software and technology*, pages 3–12.

Roweis, S. T. and Lawrence, K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, pages 2323–2326.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE T Comput*, 18(5):401–409.

Tenenbaum, J. B., Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323.

Tuzel, O., Porikli, F., and Meer, P. (2008). Pedestrian Detection via Classification on Riemannian Manifolds. *IEEE T Pattern Anal*, 30(10):1713–1727.

Walter, J. A. (2004). H-mds: a new approach for interactive visualization with multidimensional scaling in the hyperbolic space. *Inform Syst*, 29(4):273–292.