Towards Artistic Collections Navigation Tools based on Relevance Feedback

Daniele Borghesani, Costantino Grana, Rita Cucchiara

Università degli Studi di Modena e Reggio Emilia
Via Vignolese 905/b - 41100 Modena
daniele.borghesani@unimore.it, costantino.grana@unimore.it,
rita.cucchiara@unimore.it

Abstract. Artistic image collections are usually managed via textual metadata into standard content management systems. More sophisticated searches can be performed using image retrieval technologies based on visual content. Nevertheless, the problem of the information presentation remains. In this paper we try to move beyond the classic grid-styled presentation model, suggesting a novel use of relevance feedback as a navigation tool. Relevance feedback is therefore used 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 exploiting a distance based space warping on the 2D projection of the distance matrix. Multitouch gestures are employed to provide feedbacks by natural interaction with the system.

1 Introduction

The valorization of cultural heritage is probably one of the most interesting and useful applications of modern technologies of human-computer interaction and multimedia search. All the plurality of artistic masterpieces can live a complementary life through digitalization, which allows a significant reduction in management costs, an enormous expansion of public —therefore of money income— and, at the same time, a tremendous freedom of data elaboration, therefore a pleasure for the public and usefulness for experts.

One of the key aspects is the way in which information is presented, in other words how the results of a visual search, or an automatic classification based on content, is shown to users. This can constitute a significant gap between such systems and the final users, especially untrained ones. Instead an engaging user interaction design could impact positively on the success of these platforms, thus increasing the interest of people on the fruition of such artistic collection, with consequent positive spillovers on the culture, the society and the economy. Therefore, in this paper we are proposing 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. 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.

Among the different forms of art, we focused 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 (which is currently the dataset considered in this work), 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 perspectives for landscapes. For this reason, they represent a challenging dataset which allows testing the effectiveness of our proposal, not only in scientific terms (how a particular set of algorithms perform on these images) but also in "social" terms (how much interest this kind of multimedia application can gather). From now on, throughout the work, we will refer to illuminated manuscripts as the primary art collection form we are focused on.

The idea comes from the feedback we had about the project "Rerum Novarum" [1], a multimedia application we developed to enhance the fruition of artistic image collections, illuminated manuscripts in particular. Besides the combined use of visual search and relevance feedback to provide visually assisted tagging, people asked us a smart way to navigate the meaningful visual content, i.e. the pictures extracted by the illuminated pages. For this reason, we are proposing this novel interactive interface which aims at redefine the use of relevance feedback and image similarity for this kind of applications.

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 [2] or SURF [3] 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 [4]. 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 [5], 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 [6].

To solve the second problem, 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 [7], 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 [8]. An interesting proposal of this kind is the Hyperbolic-MDS by [9], 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) [10], the stochastic neighbor embedding (SNE) [11] and the local linear embedding (LLE) [12]. An older yet effective approach, especially in large scale contexts, is finally the FastMap [13] 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 Rerum Novarum: visually assisted tagging for artistic documents

Typically, the majority of the visual information retrieval systems follow the schema of Fig.1. Essentially, the system has a top-down design, and a professional effort (in terms of knowledge, documents and ontologies definition). It is necessary to provide the user with the full set of functionalities, potentially exploiting some image analysis and machine learning tools to facilitate the job. This authoritative experience of experts is used to create the annotated digital library (DL), often formalized as an ontology, that becomes the center of the application design. In this *content-centered* paradigm, the user has not got a real role of intervention inside the structure: he turns out to be a simple viewer of the retrieval results, having no real interaction with the system.

In this paper, we want to provide a more similar structure to the one in Fig.2. It is based on a *user-centered* paradigm, capable of putting together abilities, experiences and knowledge of different kinds of users, such as experts, art viewers, scholars and research communities. Instead of only assuming a static authoritative knowledge, needing a long and laborious work of visual data annotation,

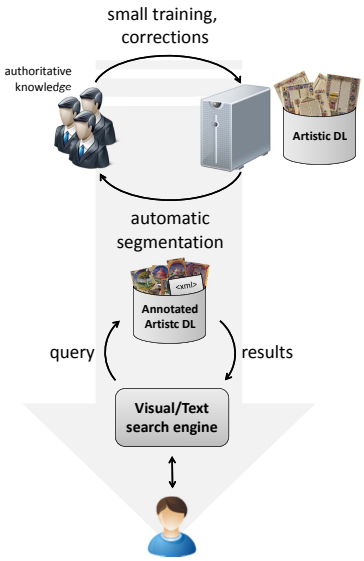

Fig. 1. The standard approach used in information retrieval systems. An expert personnel is required to include information of various nature into the system to allow the user to take advantage of proposed functionalities.

we exploit the search by visual similarity and relevance feedback to assist the process of tagging in a visual fashion, exploiting an engaging user interface.

In this context, Datta *et al.* in [15] proposed a very interesting classification of multimedia systems, based just on the user's intent, distinguishing three categories:

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

In the traditional belief, these three modalities are implemented in a separate approach, which are necessary in order to differentiate the requirements of authoritative and personal experiences. On the contrary, we propose an accommodation of all of them in the same design, allowing surfing on visual and textual contents, without excluding more general browsing functionalities neither more accurate searches. Thus, according with the schema shown in Fig.2, the normal user or expert can begin his analysis by *browsing* the pages of the documents, correcting the automatic segmentation or including a manual one if necessary. Whenever a particularly interesting detail is retrieved, the user can propose a tag to the selected picture and continue the exploration interactively

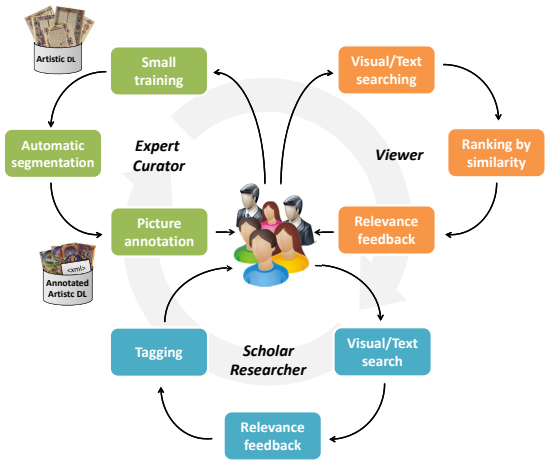

Fig. 2. Our user-centric approach. When the user is an expert, he can add his knowledge without the need of a structured representation (like tags or commentaries), and only a small subset of manually annotated training set (just some pages) is needed by the automatic segmentation [14]. When the user is a scholar or researcher, he can use the visual similarity and the relevance feedback to increase the amount of information of the system. When the user is a tourist, or an art lover, the system can be used as a simple information viewer, using the relevance feedback to improve the query results.

surfing by visual similarities. The system automatically answers with a set of similar pictures, which the user can further provide relevance feedback for. The results, marked by the user as similar, at the end may be given the same tags, so that the user will with minimal effort accomplish the otherwise demanding effort of tagging all pictures in the dataset when sharing the same visual content. In this manner the personal experience is inserted in the system by surfing and by visually assisted tagging. The tags, after a validation task if deemed necessary, become part of the collective experience: any user, art lover or researcher can so keep on analyzing the work by *searching*. Specific and reliable tags can finally give the system the possibility to filter out the visualized dataset, allowing the user to focus his attention on the sections of the work he is mainly interested on.

Basically, this is a virtuous loop in which the surfing through the artistic collection and the similarity search by visual content allows the extraction of similar pictures, i.e. pictures likely sharing the same tags; at the same time, tags will help the system to increase the embedded knowledge, and the user —while enjoying art— is facilitated on searching contents inside the artistic collection or filtering results by topic.

A very powerful analysis tool emerges by expanding this human-centered approach from a single artwork to a complete collection. The visual similarity can find meaningful results across different works. An efficient use of tags can be used to filter and organize documents and pictures across different art works. In

Fig. 3. Some examples of pictures available in this collection. Notice that we are dealing with a lot of different visual concepts: symbols with particular iconographic meanings, portraits, animals, group of people depicted in natural environment or court scenes. The pictorial style of these handmade manuscripts increases the difficulty of retrieval.

this way, the user could have literally the entire collection in his hands (see Fig.3 to get an idea of the variety of pictures available in these collections). That's the reason why we believe this approach may help in the creation of a "smart library", capable to adapt to the user's needs using efficient, yet very simple user interaction approaches.

4 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 [16] up to the product quantization [17], 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, feature space warping or machine learning approaches [18]. The first one is not efficient enough, the second one requires a full space warping (thus a full space re-encoding for ANN, and no proposals at the best of our knowledge takes into account relevance feedback in such a context); finally the learning is notoriously a heavy procedure, often requiring an offline processing and hardly capable 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. 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 [13]. 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:

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

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.

At this point, the user can 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:

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

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.

5 Interactive relevance feedback on artistic image collections

We employ this technique to improve the browsing capabilities provided in the project "Rerum Novarum" [1], a multimedia application developed to enhance the fruition of artistic image collections, showcased in ACM Multimedia 2010. The system allowed to use visual search and relevance feedback to provide visually assisted tagging. Starting from the original digitalized pages of the Holy Bible of Borso d'Este, the system performs an automatic picture segmentation using the strategy described in [14, 19, 20], possibly followed by a manual refinement. At this point, we came out with a dataset of 2282 pictures. The visual descriptor used to represent those images was the covariance matrix a simple yet effective second order statistics descriptor, which allows to embed in a very compact form a wide range of visual information: color, texture, spatial distributions and correlations between both color and edge based information.

The user was able to interact with randomly selected images, in order to start exploring the image collection. We chose to follow a different approach: the user is presented with the 2D mapping of the images and allowed to zoom and navigate (Fig. 4(a)). After identifying an interesting image (Fig. 4(b)), the

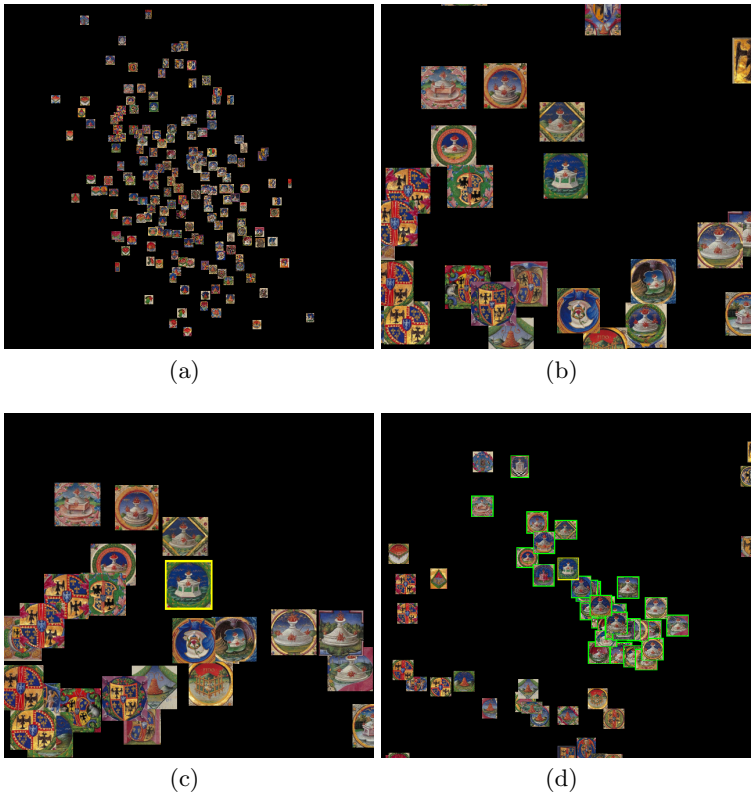

Fig. 4. Application example with the illuminated manuscript historical markings.

user selects it and the other images are rearranged to convey their distance in the feature space from the selected query (Fig. 4(c)). 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. A possible state is presented in Fig. 4(d).

6 Towards a Natural Interaction with Image Collections

The term *natural interaction* regards a human-computer interaction modality conveyed with means which are considered natural since they belong to the nature of human beings themselves [21]. The simpler and the more natural the machine interaction is, the less amount of cognitive effort is delegated to humans.

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 dynamically on the user's preferences collected continuously through relevance feedback. By exploiting a multitouch panel, the process of interacting with the system can be conveniently implemented with gesture. The removal of one or more undesired pictures is triggered with swipe gestures, while the pinch gesture allows 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.

7 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

1. Grana, C., Borghesani, D., Cucchiara, R.: Surfing on artistic documents with visually assisted tagging. In: ACM Multimed. (2010) 1343–1352
2. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. Int J Comput Vision **60**(2) (2004) 91–110
3. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF). Comput Vis Image Und **110**(3) (2008) 346–359

4. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE T Pattern Anal **27**(10) (2005) 1615–1630
5. Tuzel, O., Porikli, F., Meer, P.: Pedestrian Detection via Classification on Riemannian Manifolds. IEEE T Pattern Anal **30**(10) (2008) 1713–1727
6. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Visual Perception, Progress in Brain Research **155** (2006)
7. Heesch, D.: A survey of browsing models for content based image retrieval. Multimed Tools Appl **40** (2008) 261–284
8. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE T Comput **18**(5) (1969) 401–409
9. Walter, J.A.: H-mds: a new approach for interactive visualization with multidimensional scaling in the hyperbolic space. Inform Syst **29**(4) (2004) 273–292
10. Tenenbaum, J.B., Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science **290**(5500) (2000) 2319–2323
11. Hinton, G.E., Roweis, S.T.: Stochastic neighbor embedding. In: Neu Inf Pro Syst. (2002) 833–840
12. Roweis, S.T., Lawrence, K.: Nonlinear dimensionality reduction by locally linear embedding. Science (2000) 2323–2326
13. Faloutsos, C., Lin, K.I.: Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: ACM SIGMOD International Conference on Management of Data. (1995) 163–174
14. Grana, C., Borghesani, D., Cucchiara, R.: Automatic segmentation of digitalized historical manuscripts. Multimedia Tools and Applications (July 2010) 1–24
15. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computer Surveys **40**(2) (2008) 1–60
16. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: IEEE Symposium on Foundations of Computer Science. (2006) 459–468
17. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE T Pattern Anal **33**(1) (2011 2011) 117–128
18. Chang, Y., Kamataki, K., Chen, T.: Mean shift feature space warping for relevance feedback. In: IEEE Image Proc. (2009) 1849–1852
19. Seidenari, S., Pellacani, G., Grana, C.: Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment. British Journal of Dermatology **149**(3) (September 2003) 523–529
20. Grana, C., Borghesani, D., Cucchiara, R.: Optimized block-based connected components labeling with decision trees. IEEE Transactions on Image Processing **19**(6) (June 2010) 1596–1609
21. Baraldi, S., Del Bimbo, A., Landucci, L., Torpei, N.: Natural interaction. In: Encyclopedia of Database Systems. Springer US (2009) 1880–1885