QUEST: Towards a Multi-Modal CBIR Framework Combining Query-by-Example, Query-by-Sketch, and Text Search

Ihab Al Kabary, Ivan Giangreco, Heiko Schuldt
Department of Mathematics and Computer Science
University of Basel, Switzerland
firstname.lastname@unibas.ch

Fabrice Matulic, Moira Norrie
Institute of Information Systems
ETH Zürich, Switzerland
firstname.lastname@inf.ethz.ch

***Abstract*—The enormous increase of digital image collections urgently necessitates effective, efficient, and in particular highly flexible approaches to image retrieval. Different search paradigms such as text search, query-by-example, or query-by-sketch need to be seamlessly combined and integrated to support different information needs and to allow users to start (and subsequently refine) queries with any type of object. In this paper, we present *QUEST* (*Qu*ery by *E*xample, *S*ketch and *T*ext), a novel flexible multi-modal content-based image retrieval (CBIR) framework. *QUEST* seamlessly integrates and blends multiple modes of image retrieval, thereby accumulating the strengths of each individual mode. Moreover, it provides several implementations of the different query modes and allows users to select, combine and even superimpose the mode(s) most appropriate for each search task. The combination of search paradigms is by itself done in a very flexible way: either sequentially, where one query mode starts with the result set of the previous one (i.e., for incrementally refining or and/or extending a query) or by supporting different paradigms at the same time (e.g., creating an artificial query image by superimposing a query image with a sketch, thereby directly integrating query-by-example and query-by-sketch). We present the overall architecture of *QUEST* and the dynamic combination and integration of the query modes it supports. Furthermore, we provide first evaluation results that show the effectiveness and the gain in efficiency that can be achieved with the combination of different search modes in *QUEST*.**

I. Introduction

The widespread use of digital cameras (stand-alone or integrated in mobile phones) has led to a huge increase in the amount of images available. In order to support different search intensions, effective, efficient and in particular highly flexible approaches to image retrieval are needed. Standard text search based on (mostly manual) annotations can only be used when such metadata is available — which is more and more unlikely with growing collection sizes as the manual tagging of images is a cumbersome activity and automatic annotation of images is yet far from being an accurate process. In such cases, only content-based image retrieval (CBIR) can be applied.

In general, two fundamentally different search intensions can be identified: *known item search* and *novel item search*. Searching for a known item means that a user is looking for a seen-before image which is thus known to exist in a collection. The search task will end successfully only if the user has found this specific item. On the other hand, in novel item search, the search process successfully ends as soon as the user is provided with satisfying relevant results. For finding either known items or novel items, two main CBIR search modes are used: *query-by-example* and *query-by-sketch*. In query-by-example, a query image is required which is close enough to the user's information need and allows to obtain acceptable results. Without such a query image, it is difficult to achieve good retrieval quality, even if sophisticated relevance feedback mechanisms are available. Query-by-sketch, in turn, addresses this problem by using human generated binary sketches as query objects. This eliminates the need for finding a query image and allows users to focus on the most important details of the image(s) they are looking for. However, limited sketching skills of users are among the main challenges of query-by-sketch systems, in addition to the users' inability to correctly remember the spatial location(s) of the main object(s) within the image or their exact scale or orientation. A successful query-by-sketch system needs to tackle these aforementioned challenges to be effective. Currently available CBIR systems mainly focus on either one of the two main search modes: query-by-example or query-by-sketch – and usually combine it with keyword search. While each of these approaches is well-suited for specific search tasks, both types of systems lack support for flexibly addressing a wide range of information needs in a generic way.

In this paper, we present *QUEST* (*Qu*ery by *E*xample, *S*ketch and *T*ext), a novel framework that seamlessly combines and integrates these multiple modes of image retrieval in an effort to accumulate and flexibly combine the strengths of every individual mode. With *QUEST*, a user shall no longer be stuck with using only text search with either query-by-example or query-by-sketch techniques. Rather, she can use all these search modes in an appropriate and ad hoc defined way, tailored to her information needs. For example, a user can start a query by performing a text search using metadata to initially narrow down the search space. This can be followed by choosing a specific image from the result set for a query-by-example search. Finally, the search can be fine-tuned by superimposing the query image with such a (partial) sketch consisting of either only edges, only color, or a combination of edges and color information, to create a new, artificial query object. We introduce the architecture of *QUEST* and we present different implementations of the query modes provided. In particular, we show how the different query modes can be dynamically combined and integrated. Finally, we report on first evaluation results we have obtained that show the effectiveness and efficiency of the *QUEST* approach.

The paper is organized as follows: Section II discusses related work. Section III introduces the flexible and seamless multi-modal integration of different search techniques. Section IV provides details on the implementation of *QUEST*. Evaluation results showing the effectiveness and efficiency of *QUEST* are presented in Section V. Section VI concludes.

II. Related Work

Query-by-example systems like [1], [2], [3], [4], [5] or online search engines like Google's *similar images* search rely on the user having an initial query image that is visually similar to the image(s) being searched for. Information regarding features like color, shape, or texture can be extracted from this query image and used to find visually similar images. Feature descriptors like the Dominant Color Descriptor (DCD) [6] and the Color Layout Descriptor (CLD) [7] rely on color features, while feature descriptors such as the Scale-Invariant Feature Transform (SIFT) [8] and Speeded Up Robust Features (SURF) [9] rely on extracting scale and rotation invariant keypoints to be used in the similarity search process. On the other hand, systems that enable query-by-sketch retrieval like [10], [11], [12] do not need an initial query image close enough to the image(s) being searched for, but rather rely on the user to draw a sketch of the prominent edges as a form of initial query object. Feature descriptors such as the Edge Histogram Descriptor (EHD) [13], Angular Radial Partitioning (ARP) [14], Image Distortion Model (IDM) [15] provide different ways of storing the spatial layout of edge information. Color sketches can also be used to portray the spatial layout of colors for the required images. Several of these systems combine a couple of search modes. Integrating text search with either query-by-sketch or query-by-example is the most common. Studies [16] have shown that there is user ability and willingness to combine different search modalities in image retrieval. In this paper, we extend the integration of multiple modes of images retrieval to build on the strengths of every individual mode. The user will have the ability to dynamically select and combine these query modes in an order that seems most appropriate to the task at hand by submitting a combination of text, example images, binary sketches and/or color sketches.

III. QUEST: The Overview

In this paper, we propose a multi-modal approach to image search that allows to dynamically select and combine building blocks for gradually increasing the accuracy of retrieval results. *QUEST* attempts to improve the relevance of the results by gradually applying various search paradigms. Result staging is chosen, and the user can choose which query mode to apply in which order and thus to dynamically adjust her search strategy as needed. The user has the freedom to initiate a search process using either text input, query-by-example or query-by-sketch. This is entirely dependent on the particular search task. For example, query-by-example could be used initially in situations when an initial query image is at hand and the user wants images that are close to it; similarly, if the user wants to find similar images to discover meta information from these images, or even if searching by text initially is not returning satisfying results due to the difficulty of providing descriptive keywords or the absence of such metadata in the collection. In other cases, query-by-sketch can be initially used if there is no query image at hand that is close enough to the user's information need. Drawing binary edge and/or color sketches can then be beneficial for retrieving the required image(s). Often, an initial search leads to images that are close to the user's needs, but still do not exactly meet what the user was looking for. Hence, refining the query by applying another search mode(s) to the result set can refine the results and get the user closer in her quest for the right image. In other situations, query-by-sketch can be used as a filter to tune the results of a query-by-example search, by providing a rough sketch with the spatial layout of the prominent edges within the image, or by superimposing a sketch and a result object for dynamically creating a new query object. A user could even just input the dominant color(s), or draw a color sketch, and initiate a quick search that brings her closer to her objective.

In general, the user has the ability to narrow down the search or just re-order the results at each stage as shown in Figure 1 with every pair of consecutive result sets $R_{i+1} \subseteq R_i$, except if the user explicitly wants to avoid the filter and expand the search to cover the entire image collection. With

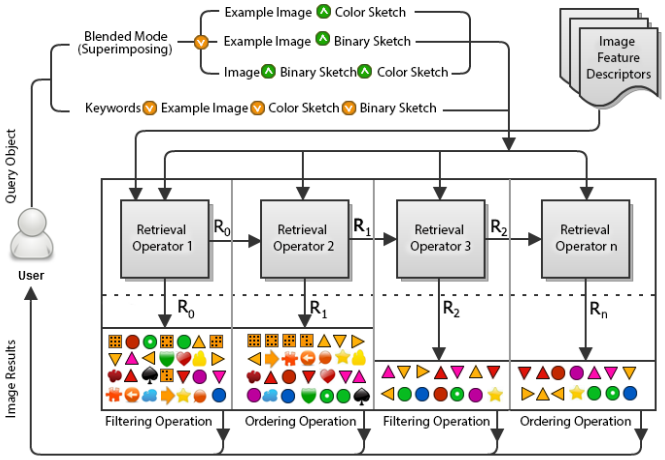


Figure 1: A *QUEST* query is incrementally refined using a variety of possible query modes.

every interaction and addition of a search mode, the results are cached to allow for easily backtracking and viewing the results before applying subsequent search modes. Results from each stage can be used as a whitelist filter for the subsequent element, or distances can be taken into account and summed up cumulatively depending on the search task. The search can be also further fine-tuned by superimposing the query image with such a (partial) sketch consisting of either only edges, only color, or a combination of edges and color information, to create a new, artificial query object.

IV. QUEST: The Building Blocks

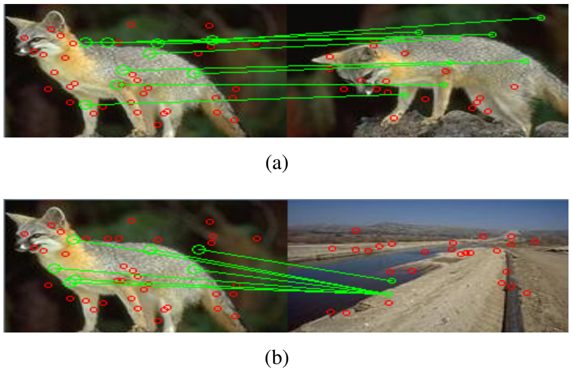
The suitability of different search modes depends on both the search task and the image collection being searched in. The following subsections present the search modes available in *QUEST* and the different building blocks implementing them. Essentially, we discuss when it is advisable to use these building blocks and how they can be combined in *QUEST*.

Text Search: When images are enriched with descriptive textual metadata, open source full text search engines can be exploited. In *QUEST* we use Lucene[1] which is an open source library with rich text search functionality sufficient for providing fast and efficient text search capabilities.

Query-by-Example: Query-by-example relies on an initial query image, which is close enough to the user's information need, in order to obtain acceptable results. Visual similarity between images can be calculated using the extracted information from the texture, shape, or color of the query image and the images in the collection. Without such a query image, it is difficult to achieve good retrieval quality. Various feature descriptors have been proposed to enable visual matching of images. SIFT [8] uses salient points found by computing the Differences of Gaussian for matching purposes. SURF [9], another interest point detector and descriptor has been proven to be multiple times faster than SIFT due to the efficient use of integral images. *QUEST* uses the JOpenSURF [17] library which we enhanced by adding multi-threading capabilities. Another enhancement has addressed the case where multiple keypoints in the query image match with a single keypoint in one of the images in the collection as shown in Figure 2. The similarity score between the query image Q and an image I from the collection is calculated without affecting the matching speed using equation 1, where Q_{kp} represents the number of matching keypoints between Q and I, and I_{kp} represent the distinct number of matched points in I.

$$SimilarityScore(Q, I) = I_{kp} + \frac{I_{kp}}{Q_{kp}} \tag{1}$$

Various CBIR systems have been built using SIFT, SURF and similar descriptors. However, in various situations, especially when searching in large image collections, relying only on these feature descriptors alone could lead to unsatisfactory results. This is due to the fact that these are local descriptors, and when used as a bag-of-words (or bag-of-keypoints) technique, they fail to capture global properties of the image. Besides, they are generally slower than lower-level feature descriptors. Hence, it is beneficial to combine these descriptors with other search modes.



(a)

(b)

Figure 2: (a) 1-to-1 matching (b) m-to-1 matching. Issue solved (no increase in matching time) in eq. 1

Query-by-Sketch: The query-by-sketch mode can be very useful in cases where an initial query image close enough to the user's information need is not available. The user can draw a binary sketch of the main contours of the image(s) being searched for, or can sketch a color representation of the image, or can use both edge and color sketches to retrieve the required image(s). Even in situations when a user has already initiated a query-by-example search, query-by-sketch could be used as a filter to further tune the results by providing a very rough sketch of the spatial layout of the edges of the desired image(s), and the results can be re-ordered accordingly.

Edge Sketches. QUEST supports the most widely used feature descriptors for sketch-based retrieval. The Edge Histogram Descriptor (EHD) [13] is frequently used for sketch-based retrieval and represents the distribution of 5 types of edges partitioned into 4x4 non-overlapping blocks. Angular Radial Partitioning (ARP) [18], [14] also uses spatial distribution of edges by means of which both sketch and image are partitioned into subregions according to an angular-radial segmentation. The image distortion model (IDM) [15] which has been used for handwritten character recognition and in medical automatic annotation tasks is yet another descriptor that has proven to work well as a distance function. It evaluates displacements of individual pixels between images within a so-called warp-range and takes patches of surrounding pixels (local context) into account for more detailed comparisons. However, it is computationally expensive when used with large warp-ranges and local context. All these algorithms tolerate a limited degree of search inaccuracies.

[1] http://lucene.apache.org

Color Sketches. To support color sketches, we combine ARP with color moments [19]. For the regions constituted by ARP, the moments are built up by the three moments mean, variance and covariance for each of the three channels in the CIE 1976 (L*, a*, b*) color space (CIELAB). A weighted, ϵ-insensitive L_1-distance is then applied to compare the moments to each other in a kNN search.

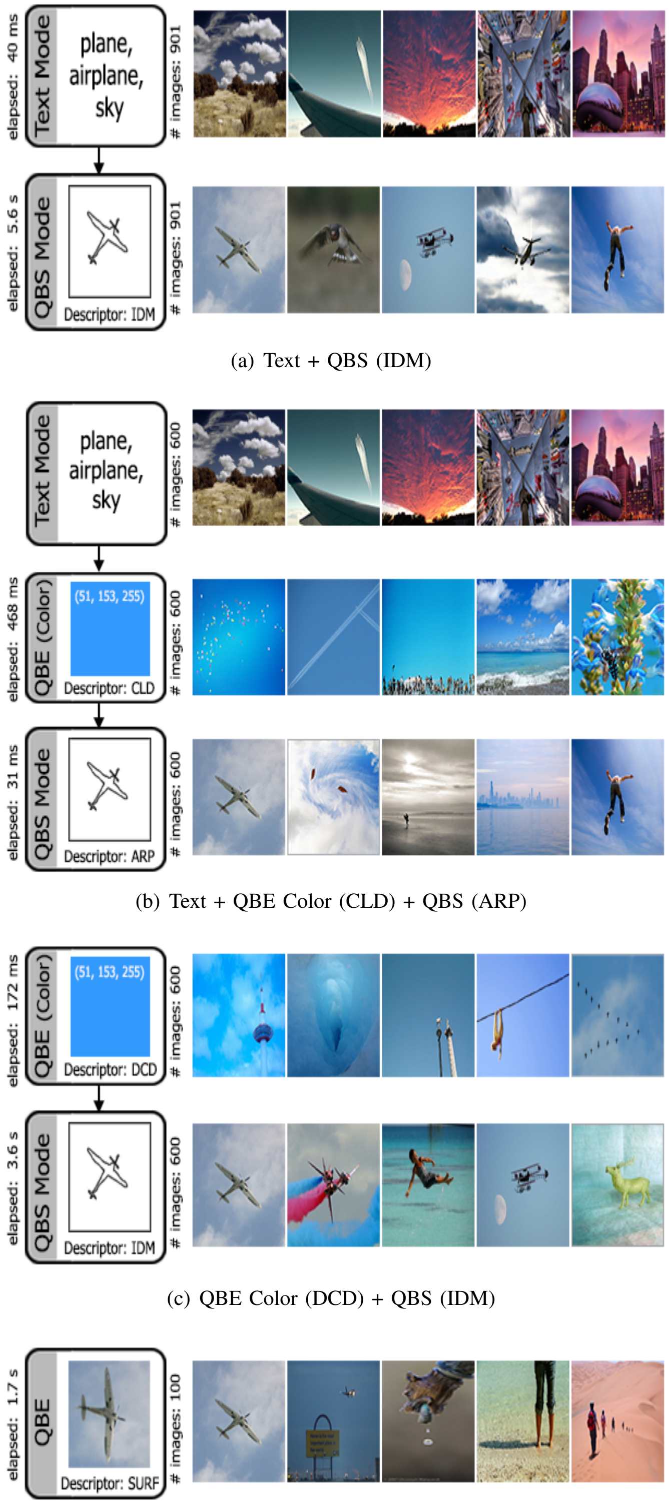
Sketch Inaccuracies and Search Invariances The user is not expected to sketch a perfect edge map representation of the image she is looking for. Limited sketching abilities of users, in addition to the users' inability to correctly remember the spatial location(s) of the main object(s) within the image, or their exact scale or orientation, are the main challenges facing the query-by-sketch search mode. In *QUEST*, we support sketch inaccuracies and provide invariances beyond the standard support of EHD, ARP, and IDM based on previous work [11] in query-by-sketch.

Color Search: Color is one of the most basic elements of an image and is used by humans to describe and distinguish images. *QUEST* supports two color search modes: the Color Layout Descriptor (CLD) [7] and the Dominant Color Descriptor (DCD) [6].

Combination of Building Blocks: All these building blocks will give the user an entire range of search modes to use, individually, in combinations, or even superimposed, depending on the query in hand. They have been designed and developed using common interfaces to allow them to be exchangeable.

V. Evaluation

In order to evaluate *QUEST* in terms of effectiveness, potential gain in efficiency, and also to assess its performance in both known item and novel item search modes, we have developed a first user interface equipped with sketch input capabilities using touch screen and interactive paper, and put the system to the test against various image collections. Initially, to test *QUEST* in the known item search mode, we used the MIRFLICKR [20] 25K collection, which is a collection of 25,000 images that score high on the flickr measure of interestingness, which takes into account how many people watched the image, commented on it, tagged it and picked it as favorite on flickr. For this collection, we picked an image of a plane (im1660.jpg) as the known item, and tried applying various search paradigms in any freely definable order to see how users could use *QUEST* to find this image. Figure 3 shows a subset of these search interactions. As shown in this figure, a user might exploit the presence of textual annotations and start searching by using keywords to decrease the search subspace, then further refine the results by using query-by-example, query-by-sketch or both in a sequential manner. If text tags are not present, initiating the search by selecting dominant color(s) could be an option, as it is extremely fast, and following this with a query-by-sketch or query-by-example would increase the the



(a) Text + QBS (IDM)

(b) Text + QBE Color (CLD) + QBS (ARP)

(c) QBE Color (DCD) + QBS (IDM)

(d) QBE (SURF)

Figure 3: Known item search (search mode combinations)

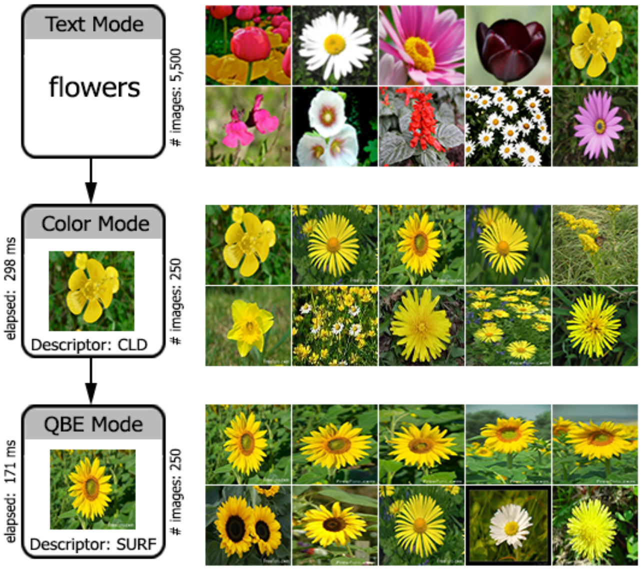


Figure 4: Novel item search (flower images)

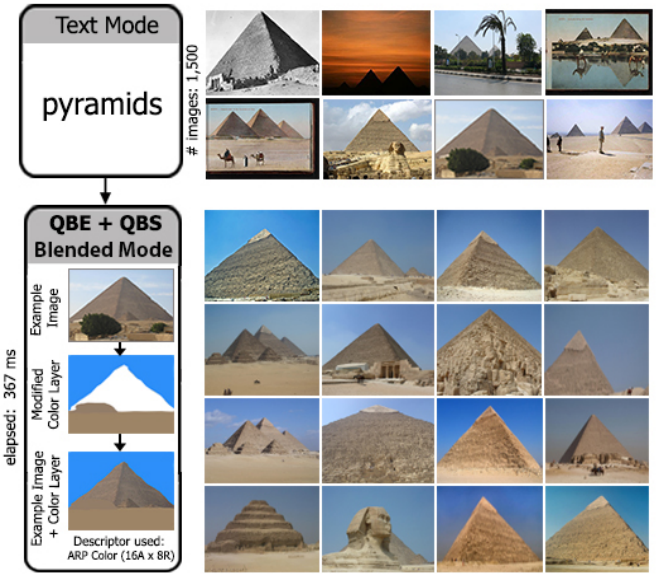


Figure 5: Blending Query-by-Example & Query-by-Sketch

possibility of finding the known item. It is also possible to start directly with a sketch, using ARP for a rough quick search or using IDM for a detailed, yet computationally more expensive search. In the case of having a close enough image, starting off with a query-by-example, using SIFT or SURF can be sufficient. Table I shows the retrieval ranks and retrieval times for text search and the various visual feature descriptors when used separately, without any pipelining. It is evident in most cases that combining various search modes is effective and gives faster response times. For example, using IDM directly with the binary sketch in Figure 3 returns the known item in the best rank possible. However, it took two and a half minutes to perform the search. On the other hand, combining IDM with an initial text and/or color search as shown in Figure 3 renders excellent results with the known item in first place also and in interactive retrieval times.

On the other hand, to test *QUEST* in the search for novel items, we had to use a different image collection, since MIRFLICKR does not contain clusters of visually similar images that would be necessary for testing novel item search. This is due to the fact that the collection is limited to 25,000 images that are highly heterogeneous in both the textual and visual description, and covers a highly diverse range of subjects. To overcome this, we have developed a simple crawler that downloads public domain images from the Google search engine. The crawler filters the search results to retrieve only "free to use or share" images. This phase acts as our text search mode. This allows us to consider very huge collections of free-to-use images and at the same time decreases the search subspace to thousands of images when using particular keywords as a textual filter. Figure 4 shows an example for a novel item search, where a user starts by searching for a flower in general (text search), then deciding on filtering the results to show yellow flowers on a green background (color search by example using CLD), finally, the user decides to find all images looking like a specific sunflower found in the results of the color search and the results renders successful results, thus, terminating the search session.

Meanwhile, Figure 5 shows how Query-by-Example and Query-by-Sketch modes can blend. The user initiates the search using a keyword, picks a close enough image, however, she does not want the green tree below the pyramid and wants a different shade of blue for the sky, thus adjusts the example image by overriding the undesired areas and therefore obtaining more satisfactory results. Finally, the main parameters used for various algorithms are displayed in Table II.

Descriptor	Retrieval Rank	Retrieval Time
Text Search	269	40 ms
CLD (QBE Color)	1189	209 ms
DCD (QBE Color)	52	172 ms
IDM (QBS)	1	151.6 seconds
ARP (QBS)	29	53 ms
SURF (QBE)	1	1.7 seconds

Table I: Retrieval rank and time for various search modes in Figure 3 separately (no combinations) emphasising the need for using search modes in combination.

Descriptor	Parameter	Value
SIFT (QbE)	Grid size	4×4
	Orientation planes	8
	Image Resolution	120 (long side)
SURF (QbE)	Hessian Octaves	5
	Hessian Octave Layers	2
	Hessian Threshold	0.0005
	Image Resolution	256 (long side)
	Point Sensitivity Ratio	0.80
IDM (QbS)	Resolution	32×32
	W. Range & L. Context	2 & 3
	Local Context	3
ARP (QbS)	Resolution	$4a \times 4r$
CLD (Color)	Resolution	8×8
DCD (Color)	Dominant Color Clusters	5

Table II: Evaluation parameters

VI. Conclusions and Outlook

We have presented *QUEST*, a framework that provides a novel and flexible approach to image retrieval by seamlessly integrating and combining different retrieval modes. *QUEST* gives the user the ability to dynamically select and combine building blocks implementing the mode(s) most appropriate for each search task. This gives a user the ability to incrementally refine a query in an intuitive manner. Furthermore, a user can use query-by-example and query-by-sketch in a blended mode by superimposing the query image with such a (partial) sketch consisting of either only edges, only color, or a combination of edges and color information, to create a new, artificial query object. We have presented first evaluation results that show the effectiveness and the gain in efficiency when combining different search modes in *QUEST*.

Starting with the first positive evaluation results we have obtained, it will be important to further evaluate all possible combinations of building blocks and to assess the framework from a usability perspective, i.e., to examine how well it will perform when deployed in a particular application context, such as publishing. The inherent flexibility and adaptability afforded by our combined search model might indeed come at the expense of increased complexity for the user, if the different query methods are not adequately managed and integrated. Hence, we will investigate mechanisms that help in suggesting the most suitable building blocks and their configuration. We intend to deploy our framework in an interactive tabletop system controlled by pen and touch input. We believe this novel interactive platform is an excellent environment to execute complex search tasks, with the pen functioning as a precision tool to perform sketch-based queries while multi-touch serves to set and manipulate the general operational context. In a final step, the pertinence of our approach will be evaluated in an extensive user study, from which we hope to derive important lessons for the design of future interactive search systems.

Acknowledgments

This work has been partly supported by the Swiss National Science Foundation, project MM-DocTable.

References

[1] M. Flickner et al. Query by Image and Video Content: The QBIC System. *Computer*, 28(9):23–32, 1995.

[2] J. Smith and S. Chang. VisualSEEk: a Fully Automated Content-Based Image Query System. In *Proceedings of MULTIMEDIA '96*, pages 87–98. ACM, 1996.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image Segmentation using Expectation Maximization and its Application to Querying, 2002.

[4] Lucia Ballerini, Xiang Li, Robert B. Fisher, and Jonathan Rees. A query-by-example content-based image retrieval system of non-melanoma skin lesions. In *Proceedings of the First MICCAI international conference on Medical Content-Based Retrieval for Clinical Decision Support*, MCBR-CDS'09, pages 31–38, Berlin, Heidelberg, 2010. Springer-Verlag.

[5] A. Arampatzis, K. Zagoris, and S. Chatzichristofis. Dynamic two-stage Image Retrieval from Large Multimedia Databases. *Information Processing and Management*, 49(1):274 – 285, 2013.

[6] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 2001.

[7] E. Kasutani and A. Yamada. The MPEG-7 Color Layout Descriptor. In *International Conference on Image Processing*, volume 1, pages 674 –677, 2001.

[8] D. Lowe. Object Recog. from Local Scale-Invariant Features. In *Proc. ICCV*, 1999.

[9] H. Bay, T. Tuytelaars, and L.Van Gool. SURF: Speeded up robust features. In *In ECCV*, pages 404–417, 2006.

[10] Y. Cao, C. Wang, Li. Zhang, and Le. Zhang. Edgel Inverted Index for Large-Scale Sketch-based Image Search. In *CVPR*, 2011.

[11] I. Al Kabary and H. Schuldt. Sketch-based Image Similarity Search with a Pen and Paper Interface. In *Proc. of International Conference on Research and Development in Information Retrieval (SIGIR)*, 2012.

[12] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. PhotoSketch: A Sketch based Image Query and Compositing System. In *SIGGRAPH*, 2009.

[13] Miroslaw Bober. MPEG-7 Visual Shape Descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):716–719, 2001.

[14] A. Chalechale, G. Naghdy, and A. Mertins. Sketch-based Image Matching Using Angular Partitioning. *IEEE Trans. SMC*, 35(1):28–41, 2005.

[15] D. Keysers, T. Deselaers, C. Gollan, and H. Ney. Deformation Models for Image Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(8):1422–1435, 2007.

[16] S. Westman, A. Lustilai, and P. Oittinen. Search strategies in multimodal image retrieval. In *Proceedings of the second international symposium on Information Interaction in Context*, IIiX '08, pages 13–20, NY, USA, 2008. ACM.

[17] C. Evans. Notes on the OpenSURF Library. Technical Report CSTR-09-001, University of Bristol, January 2009.

[18] A. Chalechale, A. Mertins, and G. Naghdy. Edge Image Description using Angular Radial Partitioning. *IEE Proc. on Vision, Image & Signal Processing*, 151(2):93–101, 2004.

[19] I. Al Kabary I. Giangreco and H. Schuldt. A User Interface for Query-by-Sketch based Image Retrieval with Color Sketches. In *Proceedings of the 34th European Conference on Information Retrieval (ECIR)*. Springer, 2012.

[20] M. Huiskes and M. Lew. The MIR Flickr Retrieval Evaluation. In *Proceedings of MIR*. ACM, 2008.