

Multimodal Relevance Feedback for Interactive Image Retrieval

Nicolae Suditu

www.idiap.ch/~nsuditu

nicolae.suditu@epfl.ch

Idiap research institute

P.O. Box 592

CH-1920 Martigny

Switzerland

PhD. Thesis Proposal

June 2010

Submitted to

École Polytechnique Fédérale de Lausanne (EPFL)

Doctoral program in Electrical Engineering (EDEE)

Thesis director

Dr. François Fleuret

Committee members

Prof. Auke Jan Ijspeert

Dr. Stéphane Marchand-Maillet



Summary

My research addresses the need for an efficient, effective and interactive access into large-scale image collections. In many cases the data of different modalities are inter-related, as for example photos and annotations in photo-sharing repositories, pictures and captions in news web-sites or x-ray scans and reports in medical databases, and I am investigating retrieval approaches that are capable of exploiting such inter-relationships (ie. multimodal information retrieval).

Most of the image retrieval approaches require an initial query before offering relevance feedback tools. The problem is that the user retrieval needs are often difficult to describe in terms of keywords and relevant images may be easily filtered out. Ferecatu and Geman proposed an innovative query-free approach. Starting from an heuristic sampling of the collection, this approach does not require any explicit query. It relies solely on an iterative relevance feedback mechanism. At each iteration, it displays a small set of images and the user is asked to show the image that best matches what he is searching for. After a few iterations, the displayed set starts to include images that satisfy the user.

My main contribution so far is an extension of the state-of-the-art approach for annotated image collections. This extension integrates indexing information extracted from both image's visual content and their associated annotation keywords. Implemented as a web-application, the approach has been evaluated by 2 groups of 30 users each, and shown to be intuitive, easy to use and efficient. For a collection of 35000 images, the approach succeeds to retrieve images that satisfy the user in less than 5 iterations in 60% of the cases. The evaluation results obtained so far motivate further investigations.

My research will remain focused on exploiting multimodal inter-relationships for increasing the retrieval capabilities of this query-free approach. A particular goal will be to adapt the relevance feedback mechanism for large-scale applications and thus one step closer to commercial applications. The starting point will be to find ways to shrink the indexing information, which will reduce the storage capacity as well as the computational effort. Another goal will be to extend and evaluate the retrieval approach for other types of multimedia (eg. medical reports, songs, movies), hopefully in collaboration with other researchers.

Keywords

multimedia information retrieval, large-scale annotated image collections, query-free interactive image retrieval, relevance feedback, multimodal, textual-based and visual content-based features, user-based evaluation.

Acknowledgements

The work of N. Suditu has been supported by the Hasler Foundation through the EMMA (ie. Enhanced Medical Multimedia data Access) project.

Introduction

Modern digital technologies produce large amounts of multimedia data that includes different modalities like images, audio, video and text. Some of the largest collections for such data are Flickr, YouTube, FaceBook, Twitter and of course the whole World Wide Web. Such amount of information creates enormous possibilities and challenges at the same time. While it is easy to get everything on-line any time, it is hard to look for anything specific.

Information retrieval needs are evolving beyond the capabilities of the straightforward text-based retrieval systems in both public and private domains. There is a noticeable need for multimedia retrieval systems that are able to provide efficient access to these large-scale multimedia collections containing millions of items. In fact, nowadays less than 1% of the World Wide Web data is in textual form, the rest being of multimedia/streaming nature.

A key characteristic of multimedia collections is that data of different modalities are inter-related as for example photos and annotations in photo-sharing collections, songs and lyrics in music collections or movies and soundtracks/transcripts in video collections. These different modalities can provide complementary indexing information that would contribute to the retrieval capabilities.

My research objective is to use inter-relationships such as the ones mentioned above and to integrate the information provided by all modalities in an unified retrieval process. Specifically, my research focuses on image collections, on combining indexing features extracted from both image's visual content and their accompanying textual information. However, the expected research results can be extrapolated to the other types of multimedia collections.

This report is organized as follows. In section 1, the research topic is defined and motivated by an overview of relevant state-of-the-art. In section 2, the preliminary research is presented along with the conducted experiments and the results obtained so far. In section 3, the research issues that will be addressed during my PhD are anticipated and linked to an estimated timetable.

1 State-of-the-art

Image retrieval, as a field of multimedia information retrieval, resides at the intersection of various disciplines such as computer vision, machine learning, information retrieval, human-computer interaction, database systems and psychology [44]. I am approaching this field from the engineering point of view, and I am trying to pay attention to aspects from other disciplines as much as possible.

My research main idea is to exploit the multimodal inter-relations of multimedia data within relevant feedback mechanisms. Specifically, my research aims to bring contributions on two main directions. Regarding the multimodal inter-relations, my research focuses on combining textual-based features with visual content-based features. Regarding the relevance feedback, my research focuses on issues related to efficiency, usability and scalability.

1.1 Image collections and indexing information

The value of a collection depends on its accessibility, which in turn depends on a corresponding relevant indexing. A decade ago, the largest image collections were the stock photography collections such as Getty Images¹ and Corbis² containing hundreds of thousand images. Images were annotated carefully with keywords from a well specified vocabulary by people with a homogeneous and professional knowledge.

Nowadays, the on-line image collections such as Flickr³ or FaceBook⁴ are orders of magnitude larger. Although many images are annotated, the keywords are less reliable due to subjective perception and less consistent due to uncontrolled vocabulary. Moreover, it is almost impossible to annotate manually all images. In order to build satisfactory retrieval applications, the indexing information should be acquired/derived automatically.

1.1.1 Textual information

The main source of textual information that is exploited by the existing search engines (eg. Yahoo!, Google) is the image's captions or the paragraphs found in the proximity of the images as they are arranged in multimedia documents (eg. web-pages with news, articles, reviews) [22]. Photo-sharing repositories and social networking web-sites support and encourage the users to provide annotation keywords for their own images and write feedback comments for the images of other users [6].

Textual-based features have been extensively investigated researched for text retrieval systems and, basically, any such features can be imported by the image

¹<http://www.gettyimages.com>

²<http://www.corbisimages.com>

³<http://www.flickr.com>

⁴<http://www.facebook.com>

retrieval systems. After the textual information is cleaned up by parsing, stemming and other techniques, the textual-based feature vectors are constructed to reflect the presence of the indexing terms [34]. My research work done so far is using textual features based on LSA (latent semantic analysis) [11]. LSA takes advantage of the implicit associations between keywords, and it escapes the unreliability, ambiguity and redundancy of individual keywords.

1.1.2 Contextual information

Modern digital equipment associates automatically images with meta-data such as date/time, location (ie. Global Positioning System coordinates), and acquisition technical details (eg. device type and configuration, resolution, luminosity, exposure settings) [10].

Photo-sharing repositories and social networking web-sites accumulate meta-data such as the number of accesses/viewings, number of references/links, popularity rankings or comments of users. These kinds of contextual information could be exploited in retrieval applications [9].

Another source of information that is currently emerging, at least in the research community, is the implicit tagging from the nonverbal behavior displayed by users while interacting with multimedia data (eg. facial expressions, vocal outbursts) [43].

1.1.3 Visual information

However, the textual and contextual information cannot fully characterize the visual content of the images. Making abstraction of its feasibility, the manual annotations are subjective and incomplete by nature. For this reason, there have been proposed the use of image processing techniques to capture automatically the visual content of images [24]. The IBM QBIC project [17] developed in 1995 is regarded as the pioneer of visual content-based retrieval systems.

The visual content, or appearance, of the images is described mathematically in vector spaces based on image processing techniques (eg. global color, texture and shape information, or a combination of these). Contrary to textual information, the visual information is abstract and does not allow for intuitive search. There is no unique explanation for the difficulties encountered in content based retrieval [37].

- The concept of “semantic gap” has been extensively used in the research community to express the discrepancy and un-correlation between the abstract vectorial representations and the actual semantic interpretation of the visual content. That is why these abstract representations were called low-level features in the first place.
- The “numerical gap” refers to the incapacity of the low-level features to characterize sufficiently the visual content of the images in order to discriminate appropriately between “relevant” and “irrelevant” images.

There are two main paradigms for using the visual content-based features. The first paradigm is to use them directly in order to define some similarity metric between images, in the same way as textual-based features are used. The second paradigm is to use them indirectly, to translate automatically the visual content-based features into textual information (eg. annotation keywords) as a pre-processing operation [21]. The main idea is to achieve automatic annotation via automatic image interpretation [12] or annotation propagation [27]. Automatic translation is a complex task, involving computer vision problems such as object recognition, and this research direction make slow progress.

Many visual content-based features have been proposed to characterize globally the color distribution, texture and edge layout and many are already included in the MPEG-7 standard [26]. My research work done so far is using visual features based on SIFT (Scale Invariant Feature Transform) [25]. SIFT feature vectors are highly distinctive and robust to affine transformations, changes in illumination and limited changes in 3D viewpoint.

1.1.4 Multimodal information

In the recent years, research confirmed that both visual content-based and textual-based features have inherent limitations, and the retrieval systems are better off if they exploit both feature types in a multimodal fashion, in order to compensate each other for their own limitations (see Smeulders et al. [37]).

The simplest approach is to simply concatenate the visual content-based and textual-based features or to combine them in other rigid manner in order to obtain *composite features*. Since the results have been encouraging [23, 14, 36], they motivated the research of more advanced combinations such as *dynamically weighted features* [47]. This research direction shows potential for progress.

1.2 Image retrieval needs and approaches

There are two extreme image retrieval needs: exploration and exploitation. The expectations of the users will always be somewhere in between, and always different. Ideally, the retrieval system should support a seamless transition between them.

- **exploration** – The user wants to browse the collection while committing to a rather vague notion of relevancy that may vary over time.
- **exploitation** – The user wants to find all the images that share some specific characteristics.

Most of the image retrieval approaches follow the standard retrieval process. A search session is initiated by submitting a query to the retrieval system. The most common type of query is a set of keywords, as in the case of text retrieval. After it retrieves the first results, some retrieval systems offer relevance feedback tools that support the user in refining the results in an iterative manner. Some retrieval systems offer more complex interfaces for tuning algorithm parameters or profile/preference parameters.

1.2.1 Query-based retrieval process

The classical image retrieval approach was to annotate each image manually based on a limited vocabulary of keywords (ie. to create manually the textual information) and, basically, to reduce image retrieval to text retrieval and to make use of the well-known and well-researched *query-by-keywords* approach [5, 41]. As in the case of text retrieval, formulating a query is more suited for the *exploitation* stage than for the *exploration* stage of the retrieval process. For *exploration*, the user must rely on his creativity to reformulate queries and to understand the indexing miscarriages.

The trend in the recent years shows that image retrieval systems must evolve beyond the capabilities of the straight-forward text-based surrogates [33]. Formulating a query might not be anymore the most efficient way of searching for images. If the annotation keywords are not fully consistent, even the most optimal query may easily exclude relevant images and include non-relevant images. Moreover, users not familiar with the keywords vocabulary will likely formulate only sub-optimal queries. All these difficulties add on top of the fact that the retrieval needs are often difficult to describe in terms of keywords.

In consequence, research proposed alternative approaches that use the visual content-based features directly in the indexing/retrieving operations. The main idea consists of specifying the query as a set of feature vectors and, then, searching the collection for the best match. The difficulty is now shifted into specifying such abstract queries, which can be done only indirectly. The most generic meanings are *query-by-visual-examples*, in which the user must provide image examples similar to what he is searching for [38], and *query-by-sketching*, in which the user must hand-draw some simple colors, textures or shapes [17]. These unconventional types of queries have their own limitations by assuming suitable image examples at hand or reasonable drawing skills [2].

1.2.2 Relevance feedback mechanisms

Another way of identifying what the user is looking for is by using relevance feedback mechanisms. In general, relevance feedback is any information about the retrieved results, given by users to a retrieval system. Whereas introduced in text retrieval [18], relevance feedback has attracted more considerable attention in the content-based image retrieval [8, 48]. Replacing the burden of formulating explicit complex queries, or having good image examples at hand, by some similarity judgments is very appealing in this new field.

One could think to make use of many sorts of information from subsequent retrieval sessions [8]. In my research, relevance feedback refers only to the information acquired in the current retrieval session by including the user in the retrieval loop [32]. For this, the session is divided into several consecutive iterations; at every round the user provides feedback regarding the retrieval results, indicating relevant images (ie. positive feedback) and sometimes also non-relevant images (ie. negative feedback). The system use this new information in order to refine the results.

While early works in MARS[4] and MindReader[19] developed mechanisms for rich feedback information (eg. ranking many images, tuning many parameters), the current consensus is that mechanisms should deal with scarce feedback (eg. marking a few “relevant” images and no tuning parameters) [7]. Obviously, the minimalist relevance feedback mechanism would require marking as “relevant” one single image at each iteration.

As reported in surveys [37, 42], there are many content-based image retrieval systems in research form but very few have been commercially developed. Scalability is crucial for an image retrieval system to be practical and realistic [46]. Some of the recently proposed pro-scalable approaches are Vimas Image Search Engine [45], Virage VIR Image Engine [1] and Cortina [31].

1.2.3 Query-free retrieval process

An innovative idea of searching images without any explicit query appeared in the work of Cox et al. [7]. The backbone of their approach was a relevance feedback mechanism based on a Bayesian framework. Fang and Geman [13] and Ferecatu and Geman [15, 16] extended the Bayesian framework and provided theoretical explanations for the main algorithms. Their work focused on using a rigid similarity metric based on low-level features extracted from the visual content of images (ie. global descriptors of color, texture and shape).

Starting from an heuristic sampling of the collection, this approach does not require any explicit query. It relies solely on an iterative relevance feedback mechanism. At each iteration, it displays a small set of images and the user is asked to show the image that best matches what he is searching for. After a few iterations, the displayed set starts to include images that satisfy the user. By hiding entirely the indexing features, the user interface is effortless and self-explanatory. Moreover, this approach is intuitively suitable to support a seamless transition between the *exploration* stage and the *exploitation* stage of the retrieval process.

1.3 Motivation to further work

The recent evolution of multimedia collections towards including inter-related modalities motivates the research of retrieval systems that are able to exploit multiple types of indexing information into a unified framework (ie. multimodal retrieval) [9, 47]. Yet, most of the existing research rely exclusively on indexing features extracted from either the image’s visual content or their accompanying annotation keywords, in an uni-modal fashion.

Moreover, only a few of the existing retrieval systems are powered specifically by relevance feedback tools. Although research agrees on their potential benefits, public image search engines provide very limited functionality of this kind [9]. More research is needed for achieving maturity in terms of efficiency, usability and scalability, which are essential characteristics for a successful system.

2 Preliminary work

My research done so far consists in an extension of an innovative retrieval approach with the major advantage of being query-free. The original retrieval approach was proposed by Ferecatu and Geman [15, 16]. Starting from an heuristic sampling of the collection, this approach does not require any explicit query. It relies solely on an iterative relevance feedback mechanism. At each iteration, it displays a small set of images and the user is asked to show the image that best matches what he is searching for. After a few iterations, the displayed set starts to include images that satisfy the user.

The original approach uses a rigid similarity metric based on low-level features extracted from the visual content of images (ie. global descriptors of color, texture and shape). My main contribution is an extension for integrating indexing features extracted from both visual content and annotation keywords of images. I propose an adaptive similarity metric that weights dynamically, at each iteration, between the visual content-based and textual-based features, depending on what the user is searching for. The effectiveness of this approach is illustrated by 2 independent user-based evaluations with 30 users each.

2.1 Retrieval process

As it was proposed in [16], the backbone of the retrieval process is based on a Bayesian framework. Having a collection $\Omega = \{1, 2, \dots, k, \dots, N\}$ of annotated images, the objective of the retrieval process is to identify the small subset $S \subset \Omega$ containing all the images that match the retrieval objectives of the user. In the retrieval process, the probabilities of relevance $p(k) = P(\{k \in S\})$ are estimated as conditional probabilities depending on the relevance feedback events.

2.1.1 Statistical framework

It is assumed that the user can decide without doubt if an image belongs to S or not. Thus, for any image $k \in \Omega$ there are two distinct possibilities, $k \in S$ or $k \notin S$, and this can be interpreted as a binary event. Naturally, S is unknown to the system and it is considered to be a random variable.

Relevance feedback is accumulated iteratively as shown in Figure 1. After the system displays a small set of images $D_t \subset \Omega$, $\|D_t\| = 8$, the user must indicate one single image $x_t^* \in D_t$ that he consider to be the most similar to S , and this relevance feedback event is denoted as $\{X_{D_t} = x_t^*\}$. Thus, the cumulative event up to iteration t can be expressed as:

$$B_t = \bigcap_{i=0}^t \{X_{D_i} = x_i^*\} \quad \forall t \geq 0 \quad (1)$$

The conditional probabilities $p_{t+1}(k) = P(k \in S | B_t)$ are estimated after each relevance feedback event. Initially, when there is no relevance feedback yet, the probabilities $p_0(k)$ are initialized with 0.5 for all $k \in \Omega$. Subsequently, assuming that the events $\{X_{D_t} = x_t^*\}$ are conditionally independent from each other given

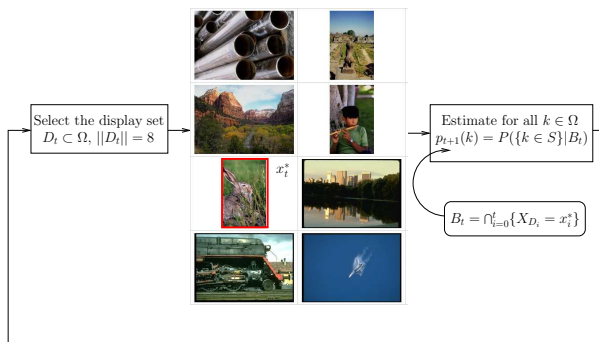


Figure 1: Relevance feedback loop. At iteration t the system displays D_t . The next iteration $t + 1$ is triggered by the relevance feedback event $\{X_{D_t} = x_t^*\}$. The system will update $p_{t+1}(k)$ for all $k \in \Omega$ and then it will select the new display set D_{t+1} .

the retrieval objectives, and using the Bayes theorem, $p_t(k)$ can be expressed recursively:

$$p_{t+1}(k) = \frac{p_t(k) \cdot P_t^+(k)}{p_t(k) \cdot P_t^+(k) + (1 - p_t(k)) \cdot P_t^-(k)} \quad (2)$$

where,

$$P_t^+(k) = P(\{X_{D_t} = x_t^*\} | \{k \in S\}) \quad (3)$$

$$P_t^-(k) = P(\{X_{D_t} = x_t^*\} | \{k \notin S\}) \quad (4)$$

One may observe that the probabilities (3, 4) enclose the user subjective perception of image similarities. My system models these probabilities based on both visual and textual features of images. I shall return to this issue in §2.2.

2.1.2 Displayed images

Sets of displayed images, namely D_t with $\|D_t\| = 8$, are generated by a clustering algorithm proposed by Fang and Geman [13]. Basically, the algorithm is growing clusters and is selecting subsequent image seeds based on the image's similarity metric (see §2.2) and their current probabilities of relevance, $p_t(k)$ for all $k \in \Omega$.

Instead of simply selecting the images with the highest probabilities of relevance, this algorithm samples the image collection with the purpose of minimizing the redundancy between the displayed images (ie. in particular, avoiding duplicates), and thus maximizing the efficiency of relevance feedback.

The initial display set D_0 is generated by running the clustering algorithm with the initial probabilities of relevance, $p_0(k) = 0.5$ for all $k \in \Omega$. The algorithm is still growing clusters but is choosing randomly between equally probable image seeds.

2.2 Extended similarity metric

For modeling the probabilities (3, 4), namely $P_t^+(k)$ and $P_t^-(k)$, I consider models of the form:

$$P_t^+(k) = \frac{\phi^+(k, x_t^*)}{\sum_{x \in D_t} \phi^+(k, x)} \quad (5)$$

$$P_t^-(k) = \frac{\phi^-(k, x_t^*)}{\sum_{x \in D_t} \phi^-(k, x)} \quad (6)$$

where ϕ^+ and ϕ^- will be designed to capture as much as possible the user subjective perception of image similarities and his decision-making behavior.

2.2.1 Joint adaptive metric

My approach uses features extracted from both visual content and annotation keywords of images. The visual-based features are derived using SIFT and the textual-based features are derived through LSA. My particular choices are specified in §2.3.2. Then, for every two images $k, l \in \Omega$, the visual-based distances $d_{visual}(k, l)$, and the textual-based distances $d_{text}(k, l)$ are obtained as Euclidean distances between the corresponding feature vectors.

Both distance types are calibrated with two monotonous functions, one for ϕ^+ and one for ϕ^- (see Figure 2). θ_{11} and θ_{21} can be viewed as saturation thresholds, above which the similarity judgments are not reliable and thus the distances are flatten out. θ_{12} and θ_{22} control the degree of coherence between the distances and the similarity judgments.

ϕ^+ and ϕ^- in (5, 6) are defined as a weighted sum of both visual-based and textual-based distances:

$$\phi^+(k, x) = \alpha \cdot \phi_{visual}^+(d_{visual}(k, x)) + (1 - \alpha) \cdot \phi_{text}^+(d_{text}(k, x)) \quad (7)$$

$$\phi^-(k, x) = \alpha \cdot \phi_{visual}^-(d_{visual}(k, x)) + (1 - \alpha) \cdot \phi_{text}^-(d_{text}(k, x)) \quad (8)$$

This is motivated by the intuition that the retrieval objectives as well as the subjective perception of image similarities are sometimes modeled better by visual features, sometimes by textual features and sometimes by a combination of both.

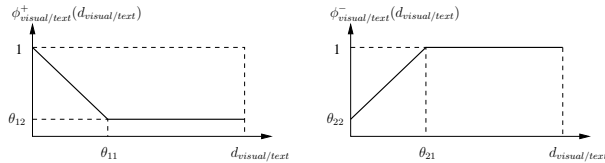


Figure 2: Calibration functions. They normalize the distances and also aim to compensate for the partial match between the distances and the user subjective perception of image similarities as explained in [15].

2.2.2 Weighting parameter

In the first iterations, both distance types are equally weighted by setting α in (7, 8) to 0.5. Subsequently, the weighting parameter α is estimated based on a Maximum Likelihood approach:

$$\alpha_{t+1}^* = \arg \max_{\alpha \in [0,1]} \frac{p_t(x_t^*)}{\sum_{x \in D_t} p_t(x)} \quad (9)$$

Immediately after the relevance feedback event $\{X_{D_t} = x_t^*\}$, before updating the probabilities $p_{t+1}(k)$ for all $k \in \Omega$, the probabilities $p_t(x)$ for the displayed images $x \in D_t$ are re-estimated for 11 discrete values of $\alpha \in \{0, 0.1, \dots, 1\}$. The optimal value α_{t+1}^* is the α that distinguishes the most x_t^* from all $x \in D_t$, and in consequence it will make the most out of the relevance feedback event $\{X_{D_t} = x_t^*\}$.

2.3 System overview

A considerable amount of effort has been invested in implementing the retrieval algorithm as a working prototype [40]. This retrieval system prototype is meant to provide a good development platform for pursuing the forthcoming research.

2.3.1 Software design

The retrieval system is designed as a web-application powered by the Apache⁵ web-server. Besides the direct advantage of permanent availability for demos and evaluations, the web-server configuration encourages the adherence to a realistic system architecture from the first research steps.

The web-application is developed in Python based on the Django⁶ framework. For optimizing the low-level critical routines, Cython was employed in translating Python into C++ and then wrapping C/C++ into Python. The application interface is shown in Figure 3.

At the heart of the system, there is a relational database based on MySQLdb⁷. Several tables store the indexing information and other meta-data (eg. location of the actual data). The pre-processing operations for extracting the image features and creating the indexing information are implemented to run efficiently in parallel processes managed by a Sun's Grid Engine (SGE⁸).

Besides the retrieval system, there is infra-structure for testing purposes. The web-application handles user accounts and stores evaluation data. Moreover, there is implemented functionality for inspecting the stored data and computing several statistics.

⁵<http://www.apache.com>

⁶<http://www.djangoproject.com>

⁷<http://www.mysql.com>

⁸<http://gridengine.sunsource.net>

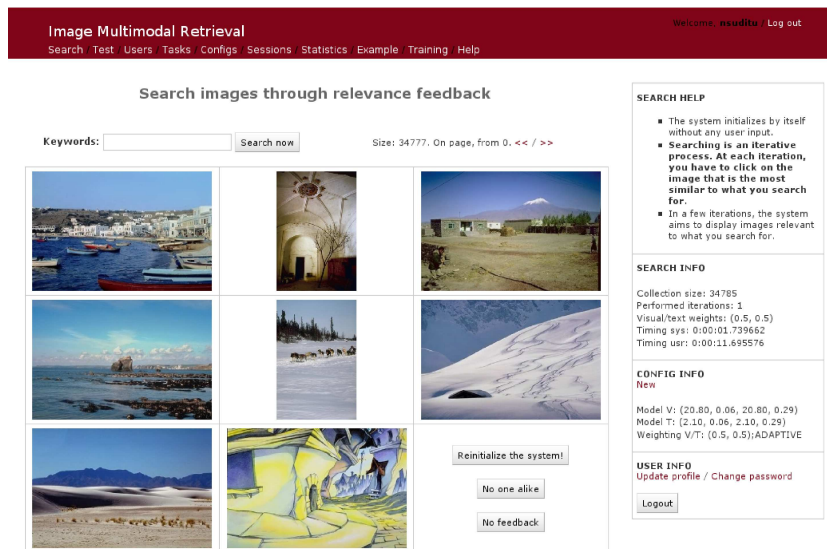


Figure 3: Web interface of the retrieval system

2.3.2 Set up details

The system was set up for a subset of the COREL image collection (ie. about 35000 photos with annotations). Each image is associated with 5-7 keywords from a vocabulary of about 5000 keywords [28, 35].

For the visual-based metric, SIFT features [25] were extracted for each image by detecting points of interest at 4 scales (ie. this resulted in 50-300 features per image). A subset of 300000 features was chosen randomly and clustered in 500 classes with the K-means algorithm. A reference SIFT vocabulary was formed by the resulted centroids. Then, a histogram-like feature vector was derived for each image by computing the membership of its own SIFT features to the centroids in the SIFT vocabulary.

For the textual-based metric, the Boolean image-keyword matrix was created by considering a vocabulary of about 5000 keywords, all the keywords that were used to annotate at least 3 images. Then, LSA was applied as explained in [11] to obtain vector representations of dimension 500.

For the calibration functions, θ_{11} and θ_{21} were chosen to saturate only after including, in average, 10% of the images in the collection. θ_{12} and θ_{22} are given the same optimum values as derived in [15], namely 0.06 and 0.29.

Computational effort required by the approach depends linearly on the collection size. Currently supporting multiple users, the web-application takes 1 second per iteration and uses 300KB cache memory per user between iterations.

2.4 User-based evaluation

Evaluation has been conducted with 2 groups of 30 users not familiar with the system. Evaluation does not rely on any apriori defined ground truth. Instead, it relies on comparing several configurations. *Bimodal-adaptive* is using both visual and textual features as described in §2.2. *Unimodal-visual* and *unimodal-textual* are particular cases obtained by setting α to 1.0 and respectively 0.0. *Pure-sampling* is a special configuration in which α is set to 0.5 and the probabilities $p_i(k)$ are fixed to 0.5 and never updated (ie. it uses the similarity metric but discards the relevance feedback). Thus, *pure-sampling* provides a fair base-line for showing the real contribution of the relevance feedback itself.

2.4.1 Evaluation scenario

Each group was assigned with 3 configurations: the first group with *bimodal-adaptive*, *unimodal-visual* and *pure-sampling*; the second group with *bimodal-adaptive*, *unimodal-visual* and *unimodal-textual*. For ensuring sufficient diversity, there were 12 semantic categories described only in words (eg. bird on water, historical site, city panorama). For ensuring comparable difficulty, they were chosen to be relevant for 1-1.5% of my collection of 35000 images.

Each user was asked to perform one searching session for each semantic category, thus 12 sessions in total. The interpretation of the semantic category in the sense of visual content was left to the user. The users were only told to end the session when they were satisfied by at least one image.

For avoiding any bias, the searching sessions were presented in a random fashion and the configurations were randomized as well. The users were not aware of which configuration was activated at a certain time. In fact, they were not aware even about the existence of these different configurations. One third of the users was available to perform 36 searching sessions in total, one for each configuration and each semantic category.

2.4.2 Results analysis

Evaluation shows that the approach is viable. All configurations using relevance feedback perform consistently better than *pure-sampling* (see Figure 4.A), and this means that the system is intuitive and able to deal with the user subjectivity in making similarity judgments.

Table 1 tells about the statistical significance of the evaluation. For each couple of configurations, I counted how many times one performed better than the other for the same user and the same semantic category, whenever there was available data. Then, I computed the binomial probabilities. In principle, a difference is statistical significant if the corresponding probability is smaller than 0.05.

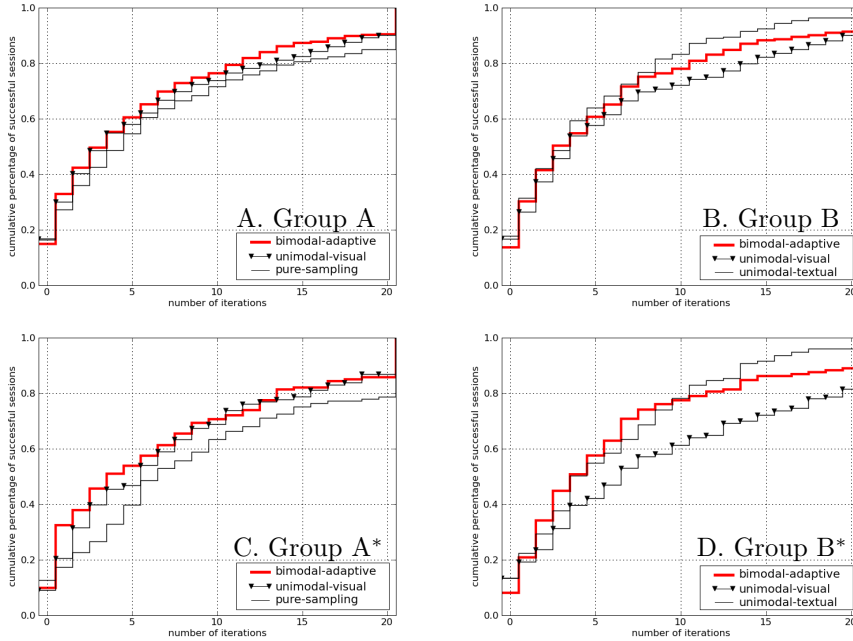


Figure 4: System performance as the cumulative percentage of successful sessions per number of iterations. Plots A-B show the average performance of the first, respectively the second group, for each of the assigned configurations. Plots C-D show the average performance of half of the corresponding group, the hardy-satisfied half.

By adding textual features, *bimodal-adaptive* and *unimodal-textual* perform significantly better than *unimodal-visual* (see Figure 4.B). *Bimodal-adaptive* and *unimodal-textual* are not significantly different (see Table 1). Since the semantic categories were specified textually, it is likely that the textual features were favored. In other contexts, visual features may become prevalent. Further evaluations should definitely address this issue.

These total averages flatten out the differences between users. One can divide the users in two sub-groups based on their performance over all configurations: easily-satisfied and hardy-satisfied users. Figures 4.C-D show how the hardy-satisfied users benefit from *bimodal-adaptive*.

Group A	Group A*	Group B	Group B*
adapt/visual (117/207) 0.025	adapt/visual (61/99) 0.007	adapt/visual (75/132) 0.048	adapt/visual (52/84) 0.010
adapt/sample (130/217) 0.002	adapt/sample (66/108) 0.008	adapt/text (72/133) 0.149	adapt/text (43/77) 0.128
visual/sample (127/210) 0.001	visual/sample (62/101) 0.007	text/visual (76/133) 0.041	text/visual (43/73) 0.050

Table 1: Binomial-test for statistical significance. (eg. for group A, *bimodal-adaptive* performed better than *unimodal-visual* in 117 times out of 207, and the probability of this to happen by chance is 0.025)

About 60% of the sessions are successfully terminated in less than 5 iterations, and 80% in less than 10 iterations. The system performance remains very reasonable when thinking of the two most extreme cases. If the collection would be arranged as a tree with 8 branches at each node, the *perfectly-structured* search will need around 2 iterations in average and $\log_8 N \approx 5$ iterations at maximum.⁹ If the collection would be totally unstructured, the *pure-random* search will need around $N/(n \cdot (L + 1)) \approx 12$ iterations in average and $N/n - \lceil L/n \rceil \gg 100$ iterations at maximum. One can remark that even the *pure-sampling* search outperforms by far the *pure-random* search.

2.4.3 Discussion

Although I did not organize an appraisal questionnaire, I received favorable informal feedback regarding the user experience. The system is unconventional but intuitive and it becomes understood in very short time, even in the first searching session. All users confirmed that the kind of similarity judgments required by the system seems natural.

Suggestions have been made to improve the user experience. For example, when users cannot make reliable similarity judgments, they would rather give *negative feedback* (ie. none of the images resembles what the user is searching for) or, at least, give *no feedback* and just ask for new images. Also, users would appreciate the possibility to *undo* the last relevance feedback iteration. Such functionalities fit well in the technique, but they were intentionally not supported in the evaluation scenario.

2.5 Conclusions

My research done so far consists in an extension of an innovative query-free image retrieval approach. Evaluation shows that exploiting the complementarity of visual content-based and textual-based features can provide better performance. Moreover, evaluation gives evidence that the approach is intuitive and able to deal with the user subjectivity in making similarity judgments. By hiding entirely the indexing features, the user interface is effortless and self-explanatory. The evaluation results give motivation for further investigations.

⁹ $N \approx 35000$, $L \approx 350$, $n = 8$ are the sizes of the image collection, semantic category and display set, respectively.

3 Research Plan

The preliminary research and the results obtained so far give me confidence to concentrate on the research direction I have chosen. The user-based evaluation helped me to identify and prioritize the most important research issues for the medium term. In this section, the research issues that will be addressed during my PhD are anticipated and linked to an estimated timetable.

3.1 Multimodal indexing information

The current retrieval system is using visual features based on SIFT (Scale Invariant Feature Transform) [25] and textual features based on LSA (latent semantic analysis) [11]. The visual-based and textual-based similarity distances are obtained as Euclidean distances between the corresponding feature vectors. Although there will be no research on new features, there will be an analysis of how the system could benefit from more advanced state-of-the-art features.

For the visual features, I will consider the generic MPEG-7 descriptors [26]. One reason is that CoPhIR image collection, for which I am planning to set up the retrieval system (see §3.3), has pre-computed and provides freely 5 global MPEG-7 descriptors for color and texture, namely the Scalable Color, Color Structure, Color Layout, Homogeneous Texture and Edge Histogram. Another reason is that the MPEG-7 standards are largely accepted and used in the research community. For the textual features, I will consider weighting methods such as tf-idf (ie. term frequency - inverse document frequency).

The current adaptive similarity metric is based on a Maximum Likelihood approach that takes in consideration only the last relevance feedback iteration. I will study and analyze other suitable approaches for combining and weighting the multimodal features according to the entire relevance feedback history.

3.2 Relevance feedback mechanism

The current retrieval system shows encouraging performance. Starting from an heuristic sampling of the collection, the system succeeds to identify what the user is searching for. Still, one can observe that the retrieval performance itself depends somehow on the collection size along with the reliability of the similarity judgments on the displayed images. Also, the computational effort required by the current approach increases linearly with the collection size. Therefore, scalability potential of the retrieval system will be investigated thoroughly considering both the functional aspect and the computational aspect.

Recalling the discussion about the general user retrieval needs in §1.1, the evaluation done so far is covering well the *exploration* stage, but it is referring only indirectly to the *exploitation* stage of the retrieval process. In fact, the current retrieval system is not well suited for the *exploitation* stage. The system performance is degrading when the probabilities of relevance become unbalanced concentrated in a small part of the collection. The clustering algorithm that

selects the displayed images is still sampling the entire collection and the relevance feedback is obviously less efficient. This effect is expected to accentuate for large collections where there are very many relevant images.

As the two stages have different requirements, Ferencat and Geman in [15, 16] suggested to invoke other more suitable retrieval approaches in the *exploitation* stage. Nevertheless, the relevance feedback mechanism should aim to support a seamless transition from one stage to another. Of course, this versatility of the relevance feedback mechanism is desirable in its own from the user interface point of view. But in my opinion, the scalability potential of the retrieval system depends closely on it as well. Therefore for the functional aspect, the starting point will be to find ways to support a seamless transition from the *exploration* stage to the *exploitation* stage of the retrieval process.

For the computational aspect, the starting point will be to find ways to shrink the indexing information, in order to reduce the storage capacity as well as the computational effort of the pre-processing operations. A breakthrough will be to find a suitable hierarchical indexing and updating strategy. This is a crucial milestone for bringing the relevance feedback mechanism up to another order of magnitude. Alternative solutions will be considered and analyzed for their impact on the system performance.

3.3 User-based evaluations

Choosing the most promising alternatives, a new retrieval system will be designed, implemented and set up for a much larger image collection. Most probably, I will set up the retrieval system for the CoPhIR image collection [3], containing images crawled from Flickr¹⁰.

CoPhIR collection contains over 100 million high-quality digital images and each image is associated in average with 5 keywords from a vocabulary of about 4000000 keywords. As reported, this collection is already used by more than 50 research institutions worldwide. The access to this collection is granted freely for research purposes through the CoPhIR Access Agreement¹¹.

As soon as the working prototype will be ready, user-based evaluations will be designed and organized in a similar fashion as the evaluation done so far. This new system will be used to evaluate the scalability beyond the technical aspects, on its impact on the user subjective relevance feedback.

3.4 Suitable applications

Alternative user interfaces will be integrated and evaluated, hopefully in collaboration with other researchers. The minimalist user interface assumed by our retrieval system seems appropriate for unconventional human-computer interactions [20]. For example, the eye-gazing trackers or real-time EEG (ie.

¹⁰<http://www.flickr.com>

¹¹<http://cophir.isti.cnr.it>

electroencephalography) neurofeedback. One application could help disabled people to communicate simple needs. Another application could let doctors to search in medical databases during surgeries.

In addition, the adaptation of the retrieval system for different applications will be explored. One of the most meaningful applications in the medical domain, where medical patient databases associates intrinsically images and reports. As clinical diagnoses benefit from comparing similar cases [29], such a retrieval system as ours could be highly appreciated.

If time allows, adaptation to other multimedia collections will be addressed as well. Movie retrieval could be an interesting application, as raw textual information can be derived from the speech transcript [39]. This textual information is of course completely different from the visual content of the underlying images, and thus the movie retrieval could benefit greatly from the multimodal approach. Songs with lyrics in music databases constitute a similar case [30].

3.5 Estimated timeline

The following timeline indicates the anticipated monthly progress of the proposed research. Some of the proposed items may be conducted in parallel.

Month 1-6 (6 months)

The scalability potential of the retrieval system will be investigated. The starting point will be to find ways to shrink the indexing information, in order to reduce the storage capacity as well as the pre-processing computational effort.

Month 3-6 (3 months)

More advanced state-of-the-art features will be considered for both the textual information and visual content of the images. There will be an analysis of how the system could benefit from them.

Month 7-12 (6 months)

Choosing the most promising alternatives, a new retrieval system will be designed, implemented and set up for a much larger image collection. This system will be used to evaluate the scalability beyond the technical aspects, on its ability to cope with the user subjective relevance feedback.

Month 13-18 (6 months)

Alternative user interfaces will be integrated and evaluated, hopefully in collaboration with other researchers. The minimalist user interface assumed by the system seems appropriate for unconventional human-computer interactions.

Month 19-22 (4 months)

Adaptation of the retrieval system for different applications (eg. medical reports) will be explored. If time allows, adaptation to other multimedia collections (eg. songs, movies) will be addressed as well.

Month 23-26 (4 months)

The last four months will be devoted to synthesize my work and write my thesis.

References

- [1] Jeffrey R. Bach, Charles Fuller, Amarnath Gupta, Arun Hampapur, Bradley Horowitz, Rich Humphrey, Ramesh C. Jain, and Chiao-Fe Shu. The Virage image search engine: An open framework for image management. In *Proceedings of Symposium on Electronic Imaging: Science and Technology – Storage and Retrieval for Still Image and Video Databases IV, IS&T/SPIE*, pages 76–87, 1996.
- [2] Kobus Barnard, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
- [3] Paolo Bolettieri, Andrea Esuli, Fabrizio Falchi, Claudio Lucchese, Raffaele Perego, Tommaso Piccioli, and Fausto Rabitti. CoPhIR: a test collection for content-based image retrieval. *CoRR*, abs/0905.4627v2, 2009.
- [4] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. BlobWorld: A system for region-based image indexing and retrieval. In *Proceedings of the 3th International Conference on Visual Information Systems*, volume 1614, page 660. Springer, January 1999.
- [5] Shi-Kuo Chang and Arding Hsu. Image information systems: Where do we go from here? *IEEE Transactions on Knowledge and Data Engineering*, 4(5):431–442, October.
- [6] Scott Counts and Eric Fellheimer. Supporting social presence through lightweight photo sharing on and off the desktop. In *Proceedings of the ACM SIGCHI international conference on Human factors in computing systems*, pages 599–606, 2004.
- [7] Ingeman J. Cox, Matthew L. Miller, Thomas P. Minka, Thomas V. Papatomas, and Peter N. Yianilos. The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.
- [8] Michel Crucianu, Marin Ferecatu, and Nozha Boujemaa. Relevance feedback for image retrieval: a short survey. In *State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction including Datamodels and Languages*, 2004.
- [9] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, April 2008.
- [10] Marc Davis, Simon King, Nathan Good, and Risto Sarvas. From context to content: leveraging context to infer media metadata. In *Proceedings of the 12th ACM international conference on Multimedia*, pages 188–195, 2004.
- [11] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, 41(6):391–407, September 1990.
- [12] Pinar Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision, ECCV - Part IV*, volume 2353, pages 349–354, January 2002.
- [13] Yuchun Fang and Donald Geman. Experiments in mental face retrieval. In *Proceedings of Audio and Video-based Biometric Person Authentication*, pages 637–646, July 2005.
- [14] Marin Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD. Thesis, INRIA-University of Versailles Saint Quentin-en-Yvelines, France, 2005.

- [15] Marin Ferecatu and Donald Geman. Interactive search for image categories by mental matching. In *Proceedings of the IEEE 11th International Conference on Computer Vision, ICCV'07*, pages 1–8, October 2007.
- [16] Marin Ferecatu and Donald Geman. A statistical framework for image category search from a mental picture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1087–1101, June 2009.
- [17] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The QBIC system. In *IEEE Computer*, volume 28, pages 23–32, September 1995.
- [18] Donna Harman. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–10, 1992.
- [19] Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos. MindReader: Querying databases through multiple examples. In *Proceedings of 24th International Conference on Very Large Data Bases*, pages 218–227, August 1998.
- [20] Alejandro Jaimes and Nicu Sebe. Multimodal human computer interaction: A survey. *Computer Vision in Human-Computer Interaction*, 3766:1–15, September 2005.
- [21] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th ACM SIGIR international conference on Research and development in information retrieval*, pages 119–126, 2003.
- [22] Mohammed L. Kherfi, Djemel Ziou, and Alan Bernardi. Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. *ACM Computing Surveys*, 36(1):35–67, 2004.
- [23] Marco La Cascia, Saratendu Sethi, and Stan Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998.
- [24] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM Transactions on Multimedia Computing, Communication and Applications*, 2(1):1–19, 2006.
- [25] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.
- [26] B. S. Manjunath, Philippe Salembier, and Thomas Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., 2002.
- [27] Florent Monay and Daniel Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings of the 11th ACM international conference on Multimedia*, pages 275–278, 2003.
- [28] Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. The truth about Corel - Evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 38–49, July 2002.
- [29] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications – Clinical benefits and future directions. *International Journal in Medical Informatics*, 73(1):1–23, 2004.
- [30] Nicola Orio. *Music information retrieval: A tutorial and review*, volume 1. Foundations and Trends in Information Retrieval, November 2006.

- [31] Till Quack, Ullrich Mönich, Lars Thiele, and B. S. Manjunath. Cortina: a system for large-scale, content-based web image retrieval. In *Proceedings of the 12th ACM international conference on Multimedia*, pages 508–511, 2004.
- [32] Yong Rui and Thomas S. Huang. A novel relevance feedback technique in image retrieval. In *Proceedings of the 7th ACM international conference on Multimedia – Part II*, pages 67–70, 1999.
- [33] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, 1999.
- [34] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
- [35] Tefko Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th ACM SIGIR international conference on Research and development in information retrieval*, pages 138–146, 1995.
- [36] Stan Sclaroff, Marco La Cascia, and Saratendu Sethi. Using textual and visual cues for content-based image retrieval from the world wide web. *Image Understanding*, 75(2):86–98, 1999.
- [37] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
- [38] John R. Smith and Shih-Fu Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the 4th ACM international conference on Multimedia*, pages 87–98, 1996.
- [39] Cees G.M. Snoek and Marcel Worring. A review on multimodal video indexing. In *Proceedings of the ICME*, volume 2, pages 21–24, 2002.
- [40] Nicolae Suditu. Image retrieval system, <http://imr.idiap.ch>. 2010.
- [41] Hideyuki Tamura and Naokazu Yokoya. Image database systems: A survey. *Pattern Recognition*, 17(1):29–43, 1984.
- [42] Remco C. Veltkamp and Mirela Tanase. Content-based image retrieval systems: A survey. *Technical Report UU-CS-2000-34, Department of Computer Science, Utrecht University, The Netherlands*, October 2000.
- [43] Alessandro Vinciarelli, Nicolae Suditu, and Maia Pantic. Implicit human - centered tagging. In *Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1–4, June/July 2009.
- [44] James Z. Wang, Nozha Boujemaa, Alberto Del Bimbo, Donald Geman, Alexander G. Hauptmann, and Jelena Tešić. Diversity in multimedia information retrieval research. In *Proceedings of the 8th ACM MIR international workshop on Multimedia information retrieval*, pages 5–12, 2006.
- [45] Yi-Leh Wu, King-Shy Goh, Beita Li, Huaxing You, and Edward Y. Chang. The anatomy of a multimodal information filter. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–471, 2003.
- [46] Atsuo Yoshitaka and Tadao Ichikawa. A survey on content-based retrieval for multimedia databases. In *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, volume 11, February 1999.
- [47] Xiang Sean Zhou and Thomas S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE MultiMedia*, 9(2):23–33, April/June 2002.
- [48] Xiang Sean Zhou and Thomas S. Huang. Relevance feedback for image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.