A Relevant Image Search Engine with Late Fusion: Mixing the Roles of Textual and Visual Descriptors*

Franco M. Segarra **Luis A. Leiva** **Roberto Paredes**
ITI – Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera, s/n - 46022 Valencia, Spain
{fsegarra,luileito,rparedes}@iti.upv.es

ABSTRACT

A fundamental problem in image retrieval is how to improve the text-based retrieval systems, which is known as "bridging the semantic gap". The reliance on visual similarity for judging semantic similarity may be problematic due to the semantic gap between low-level content and higher-level concepts. One way to overcome this problem and increase thus retrieval performance is to consider user feedback in an interactive scenario. In our approach, a user starts a query and is then presented with a set of (hopefully) relevant images; selecting from these images those which are more relevant to her. Then the system refines its results after each iteration, using late fusion methods, and allowing the user to dynamically tune the amount of textual and visual information that will be used to retrieve similar images. We describe how does our approach fit in a real-world setting, discussing also an evaluation of results.

Author Keywords

Image Retrieval, Relevant Feedback, Late Fusion

ACM Classification Keywords

H.3.3 Information Search and Retrieval: Relevance feedback, Information filtering, Retrieval models; H.5.1 Multimedia Information Systems: Evaluation/methodology

General Terms

Algorithms, Design, Human Factors, Performance

INTRODUCTION

From the very beggining of the Internet, images are a major and fast growing media. Allowing an effective search among such a huge number of online image files is a challenging task. Current approaches for retrieving relevant images have evolved from the text-based techniques used in classical information retrieval to the content-based image retrieval (CBIR) paradigm. CBIR helps to organize digital pictures by their visual content, and traditionally has involved a myriad of multidisciplinary fields such as computer vision, machine learning, human-computer interaction (HCI), database systems, statistics, and many more [1, 5].

One problem with all current approaches is the reliance on visual similarity for judging semantic similarity, which may be problematic due to the semantic gap between low-level content and higher-level concepts [4]. Relevance Feedback (RF) is a query modification scheme which attempts to capture the user's precise needs through iterative feedback and query refinement [1]. Moreover, this RF scheme can be applied using multimodal information such as visual and textual information. In this case we obtain a Multimodal Relevance Feedback system (MRF). This multimodal approach requires some fusion scheme in order to manage both modalities. In the present work we propose a late fusion approach that allows to blend the amount of textual and visual information that will be used to retrieve relevant images, overcoming thus with the inherent problem of the abovementioned semantic gap. Unfortunately, user studies of this nature have been scarce so far [1, 2]. With this demo we want to enlighten both researchers and practitioners with a straightforward methodology for developing and running a custom Relevant Image Search Engine (RISE from here onwards).

User Interaction Protocol in RISE

We implemented the probabilistic model detailed in the work of Paredes et al. [3], which is augmented with late fusion techniques. In sum, the user has in mind some relevant set of (unknown) images, and RISE's goal is to interactively discover n images of it, among the images in a fixed, finite collection of images C:

- Initially the user inputs a query q to the system.
- Then RISE provides an initial set $X_0 \in C$ of n images that are retrieved from a database according to a suitable criteria.
- These images are judged by the user, who provides a feedback by selecting which images are relevant (and, implicitly, which are not relevant).
- The system combines such feedback information with late fusion methods to obtain a new set of images X_1 depend-

*This work is supported by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).
Demo available at http://risenet.iti.upv.es/rise/

IUI'11, February 13–16, 2011, Palo Alto, California, USA.
ACM 978-1-4503-0419-1/11/02.

ing of the nature of q, i.e., text-based or visual.

- The process is repeated until the user is satisfied, i.e., all retrieved images X_i are relevant to her.

SYSTEM OVERVIEW

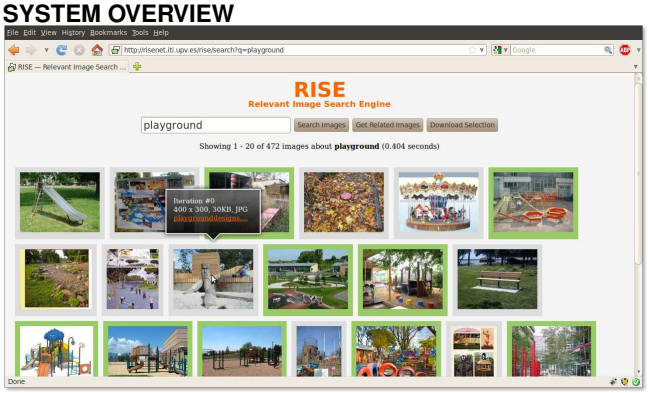

Figure 1: RISE's User Interface. After introducing a text query, the user is presented with a set of images, having to select those she finds relevant (highlighted in green). Additional information is displayed when hovering on each image.

Starting from Scratch

As starting point we used the Merriam-Webster's online dictionary to generate a list of queries to search for. After shuffling the 35,000 words in such an initial list we performed 1,500 searches for each dictionary entry among the main search engines (namely Google, Bing, and Yahoo) by using the GNU `wget` application.

Gathering Images

The image crawler stores a thumbnail as well as the feature vectors in an optimized representation for searching, requiring thus minimal storage space; e.g, one million of retrieved images take up to just 10 GB overall, including textual descriptors. The extracted signature for each image consists of a discrete distribution of region- and colour-based features.

Annotation Scheme

In order to remove the need of human assets, we automatically extract the most relevant words from the page where a certain image was included. A web page is represented by bags of weighted words, each weight being computed based on document object model (DOM) attributes. After sorting such terms by relevance (i.e., term frequency and DOM scores), we select the best 50 terms to annotate each image, which are then inserted into a MySQL database. When an annotated term is not found in the initial dictionary, it is added to the queue of the queries to be performed, enriching thus the database with new concepts.

Retrieval Procedure

To begin we use text queries to narrow the initial set of images that will be presented to the user, i.e., she types in an input field what she is looking for. Then we employ the *query-by-similar-images* paradigm with late fusion in an interactive setting (see 'User Interaction Protocol in RISE'). The best ranking R_b is computed a as linear combination of visual R_v and text R_t rankings: $R_b(\%) = \alpha R_v + (100 - \alpha) R_t$, where α accounts for the fusion percentage between visual and textual retrieval strategies. (In this way, $\alpha = 100\%$ is pure visual retrieval and $\alpha = 0\%$ is full textual retrieval.)

EVALUATION AND RESULTS

A set of 21 queries of very different nature were manually inspected. For instance, the query "tiger", where pictures of the feline can be relevant for the user, or, on the contrary, those related to the sportman Tiger Woods. We observed that, depending on the query, a pure visual strategy may help in achieving a complete set of relevant images, while in other cases pure text retrieval performs better. For that reason, we allow the user to dynamically configurable the above-mentioned parameter α. Figure 2a illustrates the gain in accuracy for the best α in comparison with both pure text and visual retrieval strategies.

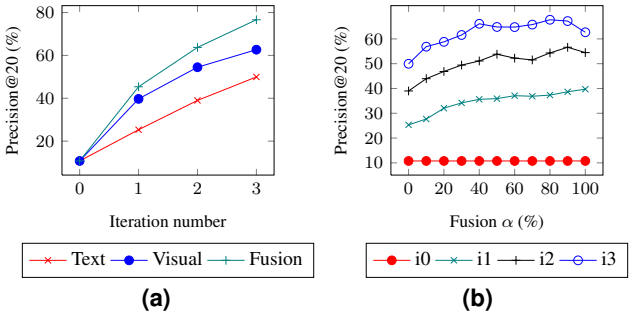

Figure 2: Experimental results. 2a: Comparison of image retrieval techniques. 2b: System precision vs. Feedback iterations.

Moreover, another remarkable result is shown in Figure 2b. It can be observed that the system response is strongly conditioned by the progression of the user iterations. That is, the more information the system has about what the user considered relevant (and non relevant), the better it can adhere to the nature of the current query. In the first iteration (i1), there is a clearly ascendent slope to the visual strategy, achieving the higher precision when full visual search is used. However, in the following iterations the best precision is not obtained on the extremes, as one might think, showing thus the importance of having a fusion method to allow seamlessly cooperate text and visual strategies.

REFERENCES

1. R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.
2. H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proc. of MIR*, pages 172–179, 2008.
3. R. Paredes, T. Deselaer, and E. Vidal. A probabilistic model for user relevance feedback on image retrieval. In *Proc. of MLMI*, pages 260–271, 2008.
4. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. PAMI*, 22(12):1349–1380, 2000.
5. J. Z. Wang, N. Boujemaa, A. D. Bimbo, D. Geman, A. G. Hauptmann, and J. Tešić. Diversity in multimedia information retrieval research. In *Proc. of MIR*, pages 5–12, 2006.