www.MMRetrieval.net: A Multimodal Search Engine

Konstantinos Zagoris Avi Arampatzis Savvas A. Chatzichristofis
Department of Electrical and Computer Engineering
Democritus University of Thrace
Xanthi, Greece
{kzagoris,avi,schatzic}@ee.duth.gr

ABSTRACT

We introduce an experimental search engine for multilingual and multimedia information, employing a holistic web interface and enabling the use of highly distributed indices. Modalities are searched in parallel, and results can be fused via several selectable methods. The engine also provides multistage retrieval, as well as a single text index baseline for comparison purposes. Initial impressions on its effectiveness are positive, while its efficiency may easily be improved.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

1. INTRODUCTION

As digital information is increasingly becoming multimodal, the days of single-language text-only retrieval are numbered. Take as an example Wikipedia where a single topic may be covered in several languages and include non-textual media such as image, sound, and video. Moreover, non-textual media may be annotated with text in several languages in a variety of metadata fields such as object caption, description, comment, and filename. Current search engines usually focus on limited numbers of modalities at a time, e.g. English text queries on English text or maybe on textual annotations of other media as well, not making use of all information available. Final rankings are usually results of fusion of individual modalities, a task which is tricky at best especially when noisy or incomplete modalities are involved.

In this paper we present the experimental multimodal search engine `http://www.mmretrieval.net` (Fig.1), which allows multimedia and multilingual queries in a single search and makes use of the total available information in a multimodal collection. All modalities are indexed separately and searched in parallel, and results can be fused with different methods depending on *a)* the noise and completeness characteristics of the modalities in a collection, and *b)* whether the user is in a need of initial precision or high recall. Beyond fusion, we also provide 2-stage retrieval by first thresholding the results obtained by secondary modalities, targeting recall, and then re-ranking the results based on the primary modality.

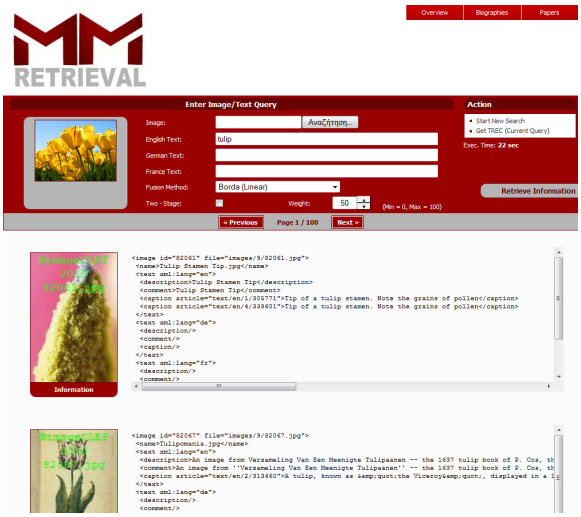

Figure 1: The www.MMRetrieval.net search engine.

The engine demonstrates the feasibility of the proposed architecture and methods on the ImageCLEF 2010 Wikipedia collection.[1] The primary modality is image, consisting of 237434 items, associated with noisy and incomplete user-supplied textual annotations and the Wikipedia articles containing the images. Associated modalities are written in any combination of English, German, French, or any other unidentified language.

2. INDEXING

To index the images, we consider the family of descriptors known as Compact Composite Descriptors (CCDs). CCDs consist of more than one visual features in a compact vector, and each descriptor is intended for a specific type of image. We index with two descriptors from the family, i.e., the Joint Composite Descriptor (JCD) [4] and the recently proposed Spatial Color Distribution (SpCD) [3]. JCD is developed for color natural images, while SpCD is considered suitable for colored graphics and artificialy generated images. Thus, we have 2 image indices.

[1] `http://www.imageclef.org/2010/wiki`

SISAP '10 September 18-19, 2010, Istanbul, Turkey.

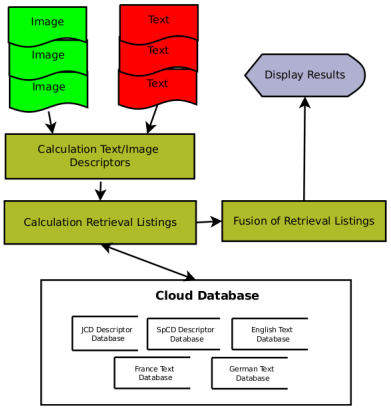

Figure 2: System's architecture.

The collection of images comes with XML metadata, consisting of a description, a comment, and multiple captions, per language (English, German, and French). Each caption is linked to the wikipedia article where the image appears in. Additionally, a raw comment is supplied which contains all the per-language comments and any other comment in an unidentified language; we do not use this field due to its great overlap with the per-language comments. Any of the above fields may be empty, noisy, or incomplete. Furthermore, a name field is supplied per image containing its filename. We do not use the supplied <license> field.

For text indexing and retrieval, we employ the Lemur Toolkit V4.11 and Indri V2.11 with the tf.idf retrieval model.[2] In order to have clean global (DF) and local statistics (TF, document length), we split the metadata per language and index them separately preserving the fields. Lemur allows searching within fields and we use this facility, as we will see below, resulting to many modalities. This, together with a separate index for the name field, results in 4 indices. Additionally, as a brute-force baseline, we also provide a single text index of all metadata and associated articles where no metadata fields or language information is used.

3. SEARCHING

The web application is developed in the C#/.NET Framework 4.0 and requires a fairly modern browser as the underlying technologies which are employed for the interface are HTML, CSS and JavaScript (AJAX). Fig.2 illustrates an overview of the architecture. The user provides image and text queries through the web interface which are dispatched in parallel to the associated databases. Retrieval results are obtained from each of the databases, fused into a single listing, and presented to the user.

Users can supply no, single, or multiple query images in a single search, resulting in $2 * i$ active image modalities, where i is the number of query images. Similarly, users can supply no text query or queries in any combination of the 3 languages, resulting in $5 * l$ active text modalities, where l is the number query languages: each supplied language results to 4 modalities, one per field described in the previous section, plus the name modality which we are matching with any language. The current beta version assumes that the user provides multilingual queries for a single search, while operationally query translation may be done automatically.

The results from each modality are fused by one of the supported methods. Fusion consists of two components: score normalization and combination. We provide two linear normalization methods, MinMax and Z-score, the ranked-based Borda Count in linear and non-linear forms, and the non-linear KIACDF. KIACDF is similar to the normalization introduced in [1], except that know-item queries are used (instead of historical) in estimating score transfer functions. We provide combination of scores across modalities with summation, multiplication, and maximum. In all fusion methods, except for where the max is used for combination, the user may select a weigh factor w, which determines the percentage contribution of the image modalities against the textual ones.

Beyond fusion, the system provides baseline searches on the single text index in two flavors: metadata only, and metadata including associated articles. In baseline searches, multilingual queries are concatenated and issued as one. Search can also be performed in a two-stage fashion. First, the text-only results of the baseline search on metadata plus articles are obtained. Then, the top-K results are re-ranked using only the image modalities which are fused by a selected method. We estimate the optimal K for maximizing the recall-oriented T9U measure, i.e. 2 gain per relevant retrieved and 1 loss per non-relevant retrieved, via the score-distributional method of [2].

4. FIRST IMPRESSIONS & OUTLOOK

In initial experiments, fusion methods using multiplication or summation seem to favor (in this order) initial precision at an expense of recall. Combination with max seems to favor recall, while two-stage retrieval seems to work best overall. Moreover, in theory, combination with max is more suitable than multiplication when descriptions are noisy or incomplete, while summation seems to provide in practice the most robust method.

We are currently planning controlled experiments in order to obtain a more concrete comparative evaluation of the effectiveness of the implemented methods. For enhancing efficiency, the multiple indices may easily be moved to different hosts.

5. REFERENCES

[1] A. Arampatzis and J. Kamps. A signal-to-noise approach to score normalization. In *Proceedings CIKM*. ACM, 2009.

[2] A. Arampatzis, J. Kamps, and S. Robertson. Where to stop reading a ranked list? Threshold optimization using truncated score distributions. In *Proceedings SIGIR*, pages 524–531. ACM, 2009.

[3] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux. SpCD - Spatial Color Distribution Descriptor - A fuzzy rule-based compact composite descriptor appropriate for hand drawn color sketches retrieval. In *Proceedings ICAART*, pages 58–63, 2010.

[4] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Proceedings SPPRA*, pages 134–140, 2009.

[2] `http://www.lemurproject.org`