# Multimedia Search with Noisy Modalities: Fusion and Multistage Retrieval

Avi Arampatzis, Savvas A. Chatzichristofis, and Konstantinos Zagoris

Department of Electrical and Computer Engineering
Democritus University of Thrace, Xanthi 67100, Greece.
{avi,schatzic,kzagoris}@ee.duth.gr

**Abstract.** We report our experiences from participating to the controlled experiment of the ImageCLEF 2010 Wikipedia Retrieval task. We built an experimental search engine which combines multilingual and multi-image search, 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 baselines for comparison purposes. Experiments show that the value added by image modalities is very small when textual annotations exist. The contribution of image modalities is larger when search is performed in a 2-stage fashion, i.e., using image search for re-ranking a smaller set of only the top results retrieved by text. Furthermore, first splitting annotations to many modalities with respect to natural language and/or type and then fusing results has the potential of achieving better effectiveness than using all textual information as a single modality. Concerning fusion, the simple method of linearly combining evidence is found to be the most robust, achieving the best effectiveness.

## 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 modalities are involved.

In this paper we present the experiments performed by Democritus University of Thrace (DUTH), Greece, in the context of our participation to the ImageCLEF 2010, Wikipedia Retrieval task.[1] The ImageCLEF 2010 Wikipedia collection has image as its primary medium, consisting of 237434 items, associated with noisy and incomplete user-supplied textual annotations and the Wikipedia articles containing the images. Associated annotations are written in any combination of English, German, French, or any

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

other unidentified language. There are 70 test topics, each one consisting of a textual and a visual part: three title fields (one per language—English, German, French), and one or more example images. The exact details of the setting of the task, e.g., research objectives, collection etc., are provided in the overview paper [7].

We built an experimental multimodal search engine, `www.mmretrieval.net` (Fig.1), which allows multiple image 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 or ranked in a 2-stage fashion. The engine demonstrates the feasibility of the proposed architecture and methods, and furthermore enables a visual inspection of the results beyond the standard TREC-style evaluation. Using the engine, we experimented with different score normalization and combination methods for fusing results, as well as 2-stage retrieval by first thresholding the results obtained by secondary modalities and then re-ranking only the top results based on fusing the primary modalities.

The rest of the paper is organized as follows. In Section 2 we describe in more detail the fusion methods we experimented with and justify their use. In Section 3 we describe the MMretrieval engine, give the details on how the Wikipedia collection is indexed and a brief overview of the search methods that the engine provides. A comparative evaluation of most implemented methods is provided in Section 4; this is based solely on additional experiments performed, since we discovered a bug affecting all our official runs involving image modalities. Conclusions are drawn in Section 5.

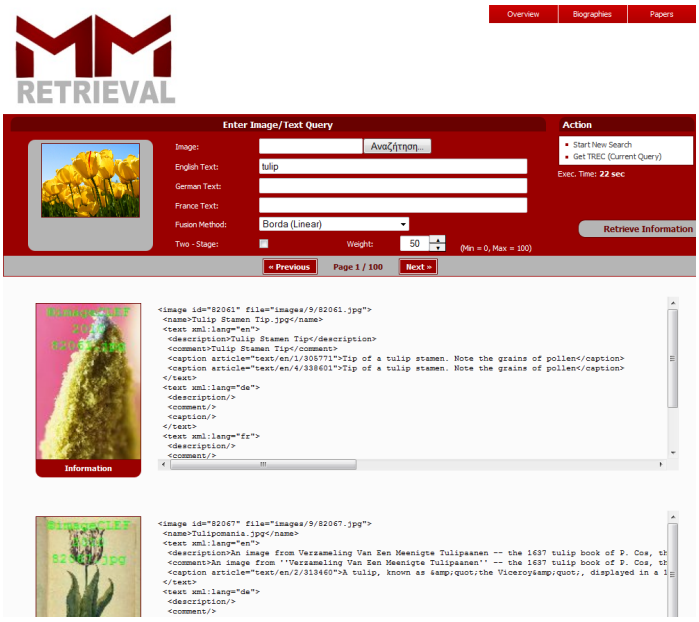


**Fig. 1.** The www.MMRetrieval.net search engine.

## 2 Fusing Modalities in IR

Let us consider a multimodal collection, i.e. a collection consisting of multiple descriptions for each of its items such as text, image, sound, etc. The term 'modality' seems to have been used interchangeably to 'medium' in the related literature. A question that arises is what an information medium really is. For example, text and image are usually seen as different media, but one can argue that text itself can come in different 'flavors' such as English, French, etc. Similarly, image information can come in different streams such as color, texture, or shape. In this respect, it might be more useful in IR to define modality as a description or representation of items in collection. For example, a multimodal collection may consist of modalities such as English text, French text, image texture, image color, etc., for each of its items.

Fusion in IR is the process of combining evidence about relevance from different sources of information, e.g. from a single modality via several retrieval models, or from several modalities. The relevance of an item to a query is usually captured in a numeric value or score. Scores for the same item across different sources may be incomparable. Thus, fusion usually consists of two components: score normalization and score combination. Assuming for now ideal scores, in the sense that are comparable across sources, we first investigate combination methods.

### 2.1 Combination

Let us assume that different sources of information produce comparable scores, e.g. probability of relevance estimates (prels), for all the items in a collection against a query. Let us investigate what the best combination of prels would be under different circumstances.

Standard fusion setups in textual IR consist of a single collection from which items are scored and ranked via several methods. Assuming that each method produces accurate prels, there is no issue of fusing the individual ranked-lists: each item would be assigned the same prel across retrieval methods, consequently using any of the methods in isolation would be sufficient. In practice, prel estimates are inaccurate with different degrees of inaccuracy, or noise, across methods. In this respect, in order to smooth out prel noise, prels for each item can be averaged or simply summed (for a constant number of sources) which has led to the popular combination method of CombSUM.

Alternatively, prel noise may be smoothed by multiplication (CombMULT), which can be seen as a proxy for their geometric average for a constant number of score sources. However, due to the fact that a single zero score would zero the multiplication, CompMULT is less robust than CombSUM. Beyond the robustness issue, we do not see a reason for preferring the geometric over the arithmetic average in the IR context. Consequently, we will not talk further about CombMULT in order to reduce the number of usable combinations.

Smoothing out noise with CombSUM implicitly assumes a non-skewed distribution of prels for an item across sources, where the average prel would make more sense. In the general case, noise can also be smoothed by taking the median prel (CombMED); this would eliminate sources with extreme noise better. Consequently, we argue that, theoretically, the most suitable combination is CombMED.

Let us now assume a multimodal collection. Having given a narrower definition of modality to mean a description or representation or a stream of information rather than medium, a standard fusion text setup mentioned above consists of a single text modality and several retrieval (or prel estimation) methods. In contrast, a multimodal setup consists of several descriptions per item as well as several suitable retrieval methods. For multimodal setups, using the same argument as above and assuming there is prel noise, a good combination method would still be CombSUM and the best theoretically would still be CombMED. However, the argument for using CombMED is now stronger as we explain next.

Consider a highly imbalanced set of modalities, e.g. one with many more text modalities than image ones. The average or sum of prels for each item would be dominated (for the good or the worse) by the textual descriptions, opening also the question of what the best balance of media is. Using the median prel does not have this problem, eliminating also the best-balance question. In summary, we argue that CombMED is the best combination method for fusing prel estimates with unknown noise characteristics, in multimodal as well as standard fusion in text retrieval. When prels are accurate or noise-free, they should be identical per item across modalities or retrieval methods so fusion is redundant.

Estimation noise can result from two sources: the retrieval model internals or the descriptions. Let us assume ideal retrieval models and focus on descriptions. Take as an example user-supplied textual annotations for images, such as caption, description, comment. These are bound to be noisy in the form of missing information, e.g. a partial description of an image or no description given at all. In this respect, prels are likely to be underestimated. Similarly, using object recognition techniques one can describe several objects (e.g. sun and beach) but usually miss more global or greater semantics (e.g. vacation), with the effect of underestimating prels. Thus, one could argue that when descriptions are missing information then the most suitable combination method is the one that takes the maximum estimated prel across modalities, i.e. CombMAX.

In practice, however, and factoring in retrieval model internals, retrieval models are bound to be more noisy in an unpredictable way at high prels. For example, a high textual score is usually a result of a low document frequency of a query word, i.e. the score is based on statistics on sparse data which are unreliable. This may also make CombMAX unreliable. Additionally, in our current setup we employ image modalities for global image features rather than local, which may sometimes capture too much rather than miss information. It is not clear whether or when prels from image modalities are under- or over-estimated, so CombMAX may not be suitable; we will further investigate this experimentally.

Similarly, when noise comes mostly in the form incomplete or empty descriptions, CombMED may have a robustness issue similar to CombMULT but not so severe: it would return zero, if more than half of the modalities return zero scores. Consequently, although we argued that theoretically CombMED is the best combination, given the noise characteristics of the collection at hand and the highly modal approach we followed by splitting the metadata and articles into several modalities, CombMED is expected to be less robust than CombSUM.

Coming back to CompSUM, the implicit assumption made is that noise has similar characteristics across modalities. If noise levels are known for the modalities, this information can be used to take a weighted average of prels instead or simply a weighed linear combination (CompWSUM); the higher the noise of a modality, the lower its weight. Appropriate weights can be estimated from training data or supplied by users, thus CompWSUM is essentially parametric.

Concluding, we arrived by argument to three usable non-parametric score combination methods for the task at hand (in a descending order of expected effectiveness): CombSUM, CombMAX, CombMED. From parametric methods, CombWSUM is the most promising and it can also become simply CombSUM for equal weights across modalities. There are many other combinations in the literature, e.g. CombMNZ, which we have not considered here.

### 2.2 Normalization

So far we have assumed ideally comparable scores across sources, e.g. in the form of probability of relevance estimates. But even most probabilistic models do not calculate the probability of relevance of items directly, but some order-preserving function of it [3]. Without training data, scores of some popular text retrieval models (e.g. tf.idf) can be turned to probabilities of relevance via the score-distributional method of [1]; however, the model does not seem to fit to the score distributions produced by our image descriptors [6].

We resort to employing two methods, which we consider them as calibration rather than normalization methods, to calibrate scores across modalities per query:

- MinMax: It maps the resulting score range linearly to the $[0, 1]$ interval.
- Zscore: A linear normalization which maps each score to the number of standard deviations it lies above or below the mean score.

MinMax does not guarantee any degree of comparability but it is a range calibrator, which may result to undesirable effects, for example: a modality with no relevant items would still return an item with a maximum score of 1. Z-score is more robust in this respect; it would calibrate high only the items separated from the heap of scores. Nevertheless, it also has its problems; it seems to assume a non-skewed distribution of scores, where the mean would be a meaningful 'neutral' score. As it is well-known, actual score distributions in text retrieval are highly skewed, clearly violating the assumption underlying Z-score. Although not very popular in IR, Z-score was used with reasonable success in [2].

As a third normalization method, we employed the non-linear Known-Item Aggregate CDF (KIACDF). KIACDF is similar to the HIS normalization introduced in [2], except that know-item queries are used (instead of historical) in estimating score transfer functions. For each modality, we issued a uniform sample of 0.5% of the collection as known-item queries, aggregated the resulting scores from all queries, and calculated their CDF. For an ad-hoc query, each resulting item score is normalized to the value of the latter CDF at the score.

## 3 www.MMRetrieval.net: A Multimodal Search Engine

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.

### 3.1 Indexing

To index 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, which we consider them as capturing orthogonal information content, i.e., the Joint Composite Descriptor (JCD) [4] and the recently proposed Spatial Color Distribution (SpCD) [5]. JCD is developed for color natural images, while SpCD is considered suitable for colored graphics and artifficially generated images. Thus, we have 2 image indices.

The collection of images at hand, i.e. the ImageCLEF 2010 Wikipedia collection, 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 or noisy. 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 in many modalities. This, together with a separate index for the name field, results in 4 indices. For English text, we enable Krovetz stemming; no stemming is done for other or unidentified languages in the current version of the system. Additionally, as a brute-force baseline, we also provide a single text index of all metadata and associated articles where no pre-processing (such as stemming) is done and no metadata fields or language information is used.

### 3.2 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

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

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.

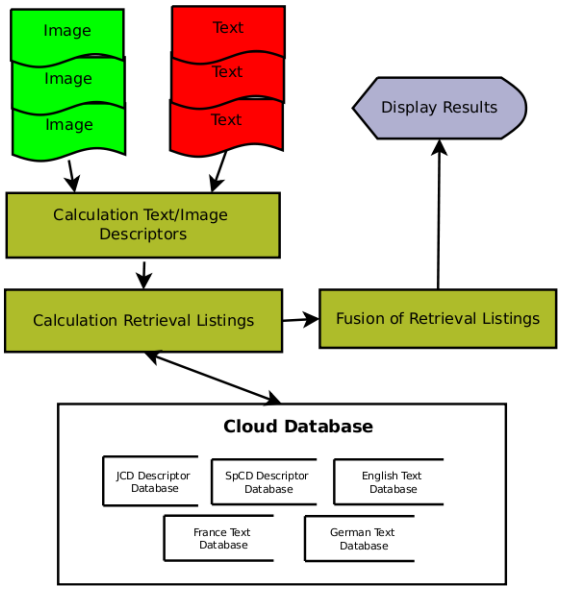


**Fig. 2.** System's architecture.

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 non-linear KIACDF, and the ranked-based BordaCount in linear and non-linear forms. Combination of scores across modalities can be done with weighted summation (CombWSUM), multiplication (CompMULT), maximum (CombMAX), or median (CombMED).

In CombWSUM, the user may select a weigh factor $w \in [0, 100]$, which determines the percentage contribution of the image modalities against the textual ones: scores from image modalities are multiplied with $w/(2i100)$ and from text modalities with $(100 - w)/(5l100)$, before summation.

For efficiency reasons, only the top-4000 results are asked from each modality. If a modality returns less than 4000 items, all non-returned items are assigned zero scores

for the modality. When a modality returns 4000 items, all non-occurring items in the top-4000 are assigned half the score of the 4000th item.

Beyond fusion, the system provides baseline searches on the single text index in two flavors: metadata only (Baseline-Metadata), and metadata including associated articles (Baseline-Any). In baseline searches, multilingual queries are concatenated and issued as one.

Search can also be performed in a 2-stage fashion. First, the text-only results of the Baseline-Any are obtained. Then, the top-$K$ results are re-ranked using only the image modalities which are fused by a selected method. By default, 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 [1].

## 4 Experiments

After the submission of our official runs, we discovered a bug in the image descriptors which affected all the runs except the text-only baselines. Fixing the bug improved effectiveness allover. Thus, we repeated all experiments with the bug-free version of the system and report them here together with additional ones. We use different run-labels than the official ones, providing more detail.

Table 1 shows that Baseline-Any performs significantly better than Baseline-Metadata. Moreover, since the associated articles constitute 3 of the modalities fused (1 per language), it also makes more sense to compare the fusion methods with Baseline-Any.

| Name | MAP | P10 | P20 | Bpref |
|---|---|---|---|---|
| MinMax CombWSUM, $w = 10$ | **0.2561** | 0.4971 | 0.4564 | 0.2997 |
| MinMax CombWSUM, $w = 20$ | 0.2272 | **0.5257** | **0.4900** | 0.2781 |
| Z-score CombWSUM, $w = 50$ | 0.1929 | 0.3714 | 0.3557 | 0.2373 |
| Z-score CombWSUM, $w = 33$ | 0.1925 | 0.3671 | 0.3457 | 0.2329 |
| Z-score CombWSUM, $w = 20$ | 0.1856 | 0.3629 | 0.3407 | 0.2267 |
| Baseline-Any | 0.1818 | 0.4243 | 0.4079 | 0.2267 |
| Z-score CombWSUM, $w = 10$ | 0.1805 | 0.3600 | 0.3371 | 0.2224 |
| MinMax CombWSUM, $w = 33$ | 0.1553 | 0.5129 | 0.4643 | 0.2104 |
| KIACDF CombMAX | 0.1550 | 0.3286 | 0.3000 | 0.1903 |
| Z-score CombMAX | 0.1539 | 0.3014 | 0.2736 | 0.1952 |
| Z-score CombWSUM, $w = 80$ | 0.1405 | 0.4243 | 0.3964 | 0.2011 |
| Baseline-Metadata | 0.1230 | 0.3700 | 0.3171 | 0.1573 |
| MinMax CombMAX | 0.1070 | 0.2914 | 0.3014 | 0.1884 |
| MinMax CombWSUM, $w = 50$ | 0.0809 | 0.4114 | 0.3364 | 0.1342 |
| MinMax CombMED | 0.0749 | 0.3414 | 0.2921 | 0.1060 |
| KIACDF CombMED | 0.0592 | 0.2957 | 0.2607 | 0.0972 |
| Z-score CombMED | 0.0255 | 0.2629 | 0.1886 | 0.0356 |
| MinMax CombWSUM, $w = 80$ | 0.0218 | 0.1771 | 0.1521 | 0.0613 |

**Table 1.** Fusion results, sorted on MAP. The best results per measure are in boldface.

### 4.1 Fusion

In Table 1 we can see that the best fusion method—and the only one beating the baseline—is the parametric CombWSUM, for both calibration methods and especially with small $w$ (i.e. small image weight). Nevertheless, its effectiveness depends largely in the choice of $w$. Fig.3 shows the impact of different choices of $w$ on MAP. For

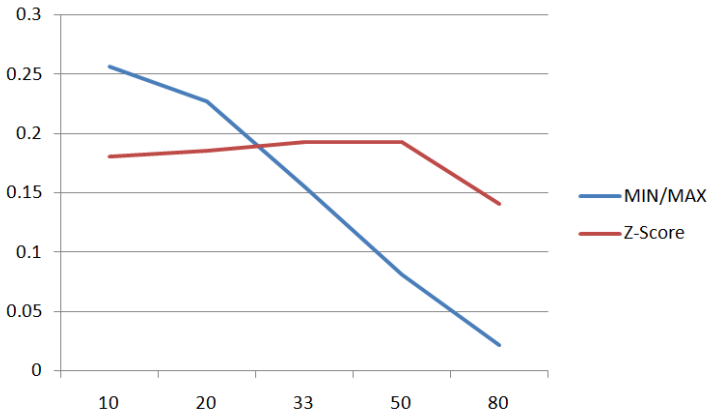


**Fig. 3.** MInMax or Z-score with CompWSUM: MAP as a function of $w$.

large $w$, CombWSUM degrades greatly with MinMax normalization. With Z-score, CombWSUM appears more robust; this is mostly a result of the highly skewed score distributions of the text modalities (in contrast to the rather symmetric ones of the image modalities [6]), producing much larger Z-scores for text than for image making CombWSUM less sensitive to $w$. We consider this as a drawback of Z-score rather than robustness. With MinMax, effectiveness keeps increasing with a decreasing $w$, which means that image modalities do not add value but rather have a negative impact.

For our 15 text and 2 image modalities (i.e. 1 query image), all modalities have roughly an equal contribution for $w \approx 12$, while the corresponding $w$ for 2 query images (i.e. 4 image modalities) is 21. Thus, for the current topic set which consists of many topics with 2 query images, CombWSUM becomes CombSUM for $w \approx 20$ or 10, suggesting that CombSUM would have been the best non-parametric method. However, we consider this accidental.

Ignoring CombSUM, the best non-parametric method is CombMAX, which is close but not beating the baseline. CombMED is disappointing due to many modalities returning zero scores; the setup has lots of noise in the form of empty (or highly incomplete) metadata, making CombMED much less robust than summation methods. We have not yet run any experiments with CombMULT, but we expect similar or worse robustness issues than CombMED.

Since image does not seem to contribute much, there is another important conclusion we can draw. First splitting the text to several indices and further to several

modalities and then retrieving by fusing results leads to improvements over using a single text index. This can be attributed to keeping cleaner keyword frequency statistics per language, separate text queries per language, as well as to using cleaner modalities conceptually.

### 4.2 2-Stage

We tested a model of 2-stage retrieval. According to this model, image modalities are considered as primary, assuming the user is searching for images with visual similarity with the query images. Since the image low-level features do not behave very well in large databases, we performed the image search in a subset. First, the text-only results of the Baseline-Any 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 [1]. The probability of relevance threshold that optimizes T9U is $\theta = 0.333$. We also tried $\theta = 0.5$ which is more precision-recall balanced and corresponds to minimizing the Error Rate.

| Name | MAP | P10 | P20 | Bpref |
|---|---|---|---|---|
| Baseline-Any | **0.1818** | **0.4243** | **0.4079** | 0.2267 |
| 2-Stage MinMax CombSUM, $\theta = 0.333$ | 0.1445 | 0.4129 | 0.3621 | 0.1978 |
| 2-Stage Z-Score CombSUM, $\theta = 0.5$ | 0.1410 | 0.3914 | 0.3586 | 0.1977 |
| 2-Stage MinMax CombSUM, $\theta = 0.5$ | 0.1401 | 0.3943 | 0.3579 | 0.1974 |
| 2-Stage Z-Score CombSUM, $\theta = 0.333$ | 0.1400 | 0.4029 | 0.3579 | 0.1934 |
| 2-Stage Z-Score CombMAX, $\theta = 0.333$ | 0.1359 | 0.3686 | 0.3200 | 0.1941 |
| 2-Stage Z-Score CombMAX, $\theta = 0.5$ | 0.1295 | 0.3657 | 0.3336 | 0.1868 |
| 2-Stage MinMax CombSUM, $\theta = 0.5$, RF | **0.1689** | 0.4043 | 0.3579 | 0.2167 |
| 2-Stage Z-Score CombSUM, $\theta = 0.5$, RF | 0.1654 | 0.4000 | 0.3579 | 0.2132 |
| 2-Stage MinMax CombSUM, $\theta = 0.333$, RF | 0.1625 | 0.3914 | 0.3579 | 0.2075 |
| 2-Stage Z-Score CombMAX, $\theta = 0.5$, RF | 0.1618 | 0.3686 | 0.3443 | 0.2095 |
| 2-Stage Z-Score CombSUM, $\theta = 0.333$, RF | 0.1614 | 0.3929 | 0.3529 | 0.2080 |
| 2-Stage Z-Score CombMAX, $\theta = 0.333$, RF | 0.1583 | 0.3800 | 0.3457 | 0.2028 |
| Baseline-Any, RF | 0.1480 | **0.4086** | **0.3686** | 0.2149 |

**Table 2.** 2-stage retrieval results, without and with RF, sorted on MAP. The best results per measure are in boldface.

Table 2 presents the results without and with relevance feedback (RF). We used pseudo/blind RF only for the first textual stage with the following parameters: top-4 items, 128 terms, and an original query weight of 0.8. These are arguably unusual RF parameters, mostly targeted to increasing the query length. It is suggested in [3] that the score-distributional method of [1] for estimating $K$ or $\theta$ works better with long queries. While the results show that long queries indeed help the method (2-stage runs with RF perform better than the RF baseline in terms of MAP), the RF baseline performs much worse than the non-RF. The choice of the RF parameters was unfortunate for the performance of the first textual phase, affecting also the 2-stage method as a whole.

Overall, the performance of the 2-stage runs with RF is competitive, achieving MAP comparable to our best non-parametric CombMAX runs. More suitable RF parameters may have led to larger improvements. Also, tighter $K$ or probability thresholds, e.g. $\theta = 0.5$, seem to work better.

### 4.3 Other Experiments

We performed a number of other experiments which we will not extensively report here, but provide only a summary.

We tried to enable RF (with the same parameters as for 2-stage) also for the fusion runs, but the results were inconsistent. While some runs improved, most of them presented lower effectiveness, pointing once more to the unfortunate choice of our RF parameters.

We also tried rank-based combinations with Borda Count in both linear and non-linear fashion, but results were far behind the score-based combinations and 2-stage; MAP was in the area of 0.06 to 0.10.

### 4.4 Further Improvements

In retrospect, it seems that we have overlooked a few things which may have led to better effectiveness:

- We should have used the raw comment field. Although most of the times it has a large overlap with the per-language comments, it sometimes carries extra text which may have been useful.
- We used stemming only for the identified English text. Stemming also the other languages could have improved effectiveness, especially when no relevance feedback is used.
- We did not experiment with the relevance feedback parameters at all, but used rather unusual values which, based on previous literature, we assumed they would improve the quality of $K$ in 2-stage runs. Although we achieved the target, other parameter values may have led to better effectiveness overall.
- $K$ is estimated to optimize a certain evaluation measure. We have tried two arbitrary measures, T9U and Error Rate. A suitable measure should be tight to the expected effectiveness of image search, a venue we have not explored.

For enhancing efficiency, the multiple indices should (and can easily) be moved to different hosts.

## 5 Conclusions

We reported our experiences from participating to the controlled experiment of the ImageCLEF 2010 Wikipedia Retrieval task. We built an experimental search engine which combines multilingual and multi-image search, 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 2-stage retrieval, as well as a single text index baselines for comparison purposes.

After various experiments, we arrived to a conclusion others had drawn before: the value added by image modalities is very small (or even negative) when textual annotations exist. This suggests that image retrieval is a problem which is far from being solved. In a more positive note, the contribution of image modalities is positive when search is performed in a 2-stage fashion, i.e., using image search for re-ranking a smaller set of only the top results retrieved by text. All these suggest that image retrieval can be reasonably effective in small databases, but it does not scale up well.

Focusing on text, first splitting annotations to many modalities with respect to natural language and/or type and then fusing results has the potential of achieving better effectiveness than using all textual information as a single modality. We attributed this to having cleaner keyword frequency statistics and separate text queries per language, as well as to using cleaner modalities conceptually.

Concerning fusion, the simple method of linearly combining evidence is found to be the most robust, achieving the best effectiveness. Combining by taking the max score across modalities is also competitive. Fusion is greatly affected by the degree of comparability of the scores combined. We tried two score calibration methods, Z-score and MinMax, and the latter achieved the best results. Nevertheless, whether any of the methods employed achieves comparability is questionable (Sec.2.2); we consider score normalization an open problem which its solution has the potential to greatly improve fusion, as well as result-merging in distributed retrieval.

All in all, we are satisfied with our results as first-time participants. Our best MAP result of 0.2561 (achieved post-submission with the bug-free version of the engine) would have ranked as the second-best run among the runs of all other participants. Moreover, we have experimentally identified some good methods for dealing with such tasks and directions for further improvements (Sec.4.4).

## References


1. Arampatzis, A., Kamps, J., Robertson, S.: Where to stop reading a ranked list? Threshold optimization using truncated score distributions. In: Proceedings SIGIR. pp. 524–531. ACM (2009)
2. Arampatzis, A., Kamps, J.: A signal-to-noise approach to score normalization. In: Proceedings CIKM. ACM (2009)
3. Arampatzis, A., Robertson, S., Kamps, J.: Score distributions in information retrieval. In: ICTIR. Lecture Notes in Computer Science, vol. 5766, pp. 139–151. Springer (2009)
4. Chatzichristofis, S.A., Boutalis, Y.S., Lux, M.: Selection of the proper compact composite descriptor for improving content based image retrieval. In: Proceedings SPPRA. pp. 134–140 (2009)
5. Chatzichristofis, S.A., Boutalis, Y.S., Lux, M.: SpCD - Spatial Color Distribution Descriptor - A fuzzy rule-based compact composite descriptor appropriate for hand drawn color sketches retrieval. In: Proceedings ICAART. pp. 58–63 (2010)
6. Chatzichristofis, S.A., Arampatzis, A.: Late fusion of compact composite descriptorsfor retrieval from heterogeneous image databases. In: SIGIR. pp. 825–826. ACM (2010)
7. Popescu, A., Tsikrika, T., Kludas, J.: Overview of the wikipedia retrieval task at imageclef 2010. In: Working Notes of CLEF 2010, Padova, Italy (2010)