Multi-modal image search for large-scale applications

Petra Budikova
Faculty of Informatics,
Masaryk University, Brno,
Czech Republic
budikova@fi.muni.cz

Michal Batko
Faculty of Informatics,
Masaryk University, Brno,
Czech Republic
batko@fi.muni.cz

Pavel Zezula
Faculty of Informatics,
Masaryk University, Brno,
Czech Republic
zezula@fi.muni.cz

ABSTRACT

One of the major challenges of data processing is the efficient multi-modal searching in large collections of complex data objects. In this paper, we present first results of a complex study that focuses on a comparison of various image retrieval methods in terms of effectiveness, efficiency, and scalability. We introduce a classification of possible approaches, outline the methodology used for evaluation, and present results of our first experiments.

1. INTRODUCTION

With the rapid growth of both volume and diversity of digital data, traditional attribute-based or text-based retrieval is no longer satisfactory for answering users' information needs. Complex data objects such as multimedia need to be managed and searched in a number of applications. For the particular subtask of image retrieval, the use cases range from scientific data management to entertainment, security, and surveillance.

First attempts at image searching followed the established lines of the attribute-based and text retrieval, and organized images with respect to their descriptive metadata. However, this solution is not applicable in a number of cases, as the metadata is often not available. A more recent content-based approach is more general, as it utilizes inherent features of the data objects. The content-based retrieval represents a whole class of approaches that exploit various characteristics of the images, such as global image features (e.g. MPEG7 color, shape, or texture descriptors [11]), local image features (e.g. SIFT [9]), face recognition, etc.

Recent research indicates that in general, it is not likely to achieve satisfactory results by applying retrieval methods that exploit only one modality, i.e. one projection of the complex object into the reduced feature space used for data management. This is caused by several reasons: 1) each modality only reflects a specific perspective of the complex object, which may not agree with the actual users' subjective view (the *semantic gap* problem); 2) as indicated above, a particular modality may not be applicable in some situations; 3) in large-scale applications, a single modality is typically not distinctive enough to distinguish relevant objects from the irrelevant ones. Therefore, latest data management techniques focus on a *multi-modal retrieval* that combines multiple orthogonal views on objects [3, 5, 12].

Already, a number of solutions for multi-modal image retrieval exist and are rapidly developing. Different research communities with different backgrounds are pursuing various research directions. To mention two examples from the opposite ends of the spectrum, there is the practical Google search based on text retrieval with content-based *reranking* [6], and the theoretical Threshold algorithm for *fusion* of multiple single-modal search results [4]. Unfortunately, no thorough comparisons of performance in terms of response times, scalability and retrieval precision are available due to the longstanding problem of image search benchmarking [8].

2. OBJECTIVES

In our research, we focus on the large-scale, interactive image retrieval. In this context, one of the most important qualities of the retrieval process is its speed and scalability. At the same time, it is desirable to support multi-modal searching as the current results indicate that this is a promising way to effective management of large data. However, majority of existing solutions were proposed for specialized, small-scale applications, and have considerable drawbacks when applied on voluminous data. Therefore, our objective is to determine which techniques are suitable for large-scale multi-modal image retrieval, and identify the factors that influence the performance of individual methods.

To discover which search methods show potential for the large-scale retrieval, we perform an extensive research of the existing approaches, analyze their properties and select those applicable to large-scale retrieval. These are then subject to a thorough evaluation of both the search costs and the relevance of results. While the costs can be measured rather precisely and, to a certain degree, even modeled theoretically, the relevance can only be assessed by user-satisfaction experiments. To achieve this, we created a novel evaluation platform, exploiting our framework for content-based searching [1] and a large collection of real-world image data. In the experimental evaluation, we focus on the popular modalities of image retrieval – global image descriptors, local image descriptors, and text annotations [3]. However, we are also interested in different pseudo-relevance feedback methods which can be utilized to refine the search results.

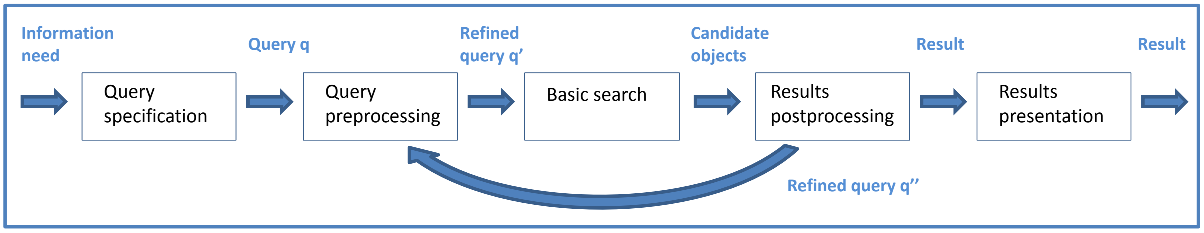

Figure 1: Basic phases of the retrieval process.

In this work, we present the results of a first phase of our research, which focuses on the effective and efficient processing of large image collections with high-quality text metadata. A typical source of such collection is a photostock web site, where the authors upload their photos along with rich and precise annotations to improve their findability. In such crowd-sourcing manner, large amounts of highly informative data are created, which can be further exploited for information mining. In this work, we consider two application scenarios: image search with a multimodal query, and automatic image annotation. The main objectives of this paper are the following:

- Classification of approaches to multi-modal retrieval: In Section 3, we present a model of a search engine and identify components that are most important for the overall performance. Next, we focus on the individual search phases, discuss existing approaches, and select significant representants for experimental evaluation.
- Comparison of different approaches to multi-modal retrieval over a large-scale, real-world dataset: Using a novel evaluation platform described in Section 4, we study the effectiveness and efficiency of various query processing methods in Section 5. In particular, we debate the individual fusion scenarios, the influence of pseudo-relevance feedback, and the variance of performance between different types of queries. Section 6 summarizes the results and outlines future research.

3. MULTIMODAL IMAGE RETRIEVAL

Image retrieval is a complex task with a number of subproblems. In this section, we first specify our target issues in the context of the whole search process. Then, we focus on the individual approaches and select methods for the experimental comparison. Due to space restrictions, we are only able to mention the most significant related work.

3.1 Search model

Figure 1 presents a simple model of a search engine, identifying several basic components of the retrieval process. The key component of each retrieval system is the central *basic search* phase, which represents the actual identification of candidate objects over the whole dataset. The surrounding *query preprocessing* and *results postprocessing* phases comprise additional methods that may be used to improve the retrieval performance but are not applied on all data objects. These three phases together may be evaluated repeatedly in a relevance-feedback loop. The first and the last phase in the model cover supplementary methods that may be used to formulate the query and to present the results to user, respectively. The model attempts to describe all steps that may be part of the retrieval process but not necessarily all search engines implement them all.

In terms of results quality, the overall performance of the search system is determined mainly by the combined effectiveness of the three central phases. Out of these, the basic search and results postprocessing are more tightly related, while the query preprocessing can be considered rather independent. Furthermore, the preprocessing is often more the subject of a domain understanding than data management. For the overall costs, the basic search phase efficiency is by far the most crucial one since it is the phase where the bulk of the data needs to be processed. To reduce the costs, approximate basic search with result-refinement postprocessing is often applied. Thus, the basic search and eventual postprocessing form the core of each search system.

In our study, we focus on determining efficient and effective techniques for these two phases, assuming a query defined either by a visual example (for the image annotation application), or a text and visual component (multi-modal image retrieval). In both cases, the query is evaluated over text-and-visual data. In the following sections, we show how different modalities derived from such data can be exploited in the basic search and postprocessing phases.

3.2 Basic search

To transform the input query object into a set of candidate objects, the basic search techniques need to survey the whole dataset. In large-scale applications, a lot of attention is devoted to the methods of efficient data indexing which allow fast retrieval of the candidates. In case of multi-modal searching, another design decision needs to be taken considering the involvement of the individual modalities in the search process. A number of solutions are based on a sequential integration of modalities, utilizing a single modality in the basic search and other modalities in the following phase. Alternatively, a multi-modal basic search can be executed when a multi-modal query is provided. In the following sections, we discuss both single-modal and multi-modal basic search.

3.2.1 *Single-modal basic search*

The single-modality retrieval is well known to the database community. Depending on the type of the modality, different data organization tools can be applied. Traditional relational databases are used for attribute data, whereas specialized index structures have been developed for text retrieval [10]. The content-based searching is of a more recent origin but mature solutions exist already, exploiting the query-by-example paradigm and similarity-based data organization [13].

3.2.2 Multi-modal basic search

Devising multimodal basic search methods is a very important and lively field of contemporary research, often denoted as *information fusion* [7, 14]. We can distinguish several types of the fusion, differing in the manner of combination and the level of involvement of the individual modalities: early fusion, late fusion, and inherent fusion. Figure 2 depicts the various approaches applied on the text-and-visual image search example.

Early fusion Early fusion search methods work throughout the retrieval process with complex data objects which contain multimodal information. This approach is also denoted as the *joint features model.* Its strong advantage is that maximum information is available for each data object which can be utilized to obtain additional knowledge, typically exploiting relations between individual modalities. Using all available information, a similarity function is proposed and employed in creating a search structure that is used for data management. While this straightforward solution can provide a high-quality searching, it also has some serious disadvantages. The data objects and similarity function are often complex, thus the computation is costly in both time and storage space. In addition, the early fusion approach does not allow any flexibility in the combination of modalities (e.g. the weights for the individual modalities) as it is necessary to have a pre-built index with all its settings fixed.

The simplest early fusion method rests in concatenating individual feature representations, as reported e.g. in [14] for image and text. Most of the research in this area is focused on the mining of semantic relations between the modalities. It is worth noticing that the early fusion approach in fact transforms the multi-modal search into a single-modal retrieval with one complex modality.

Late fusion As pointed out in [14], late fusion is the most frequently used technique in image-text fusion. Using an independent search system for each modality, a ranked list of the most relevant objects with respect to the given modality is retrieved. Next, these results are merged to form the final result. This solution enables to use any number of existing systems that work over the same data, and combine their results flexibly according to the specific application needs. In addition, the single-modal searches can be run in parallel and only the aggregation phase needs to be centralized. The late fusion solution requires no specialized preprocessing and allows real-time setting of parameters. On the other hand, the aggregation phase may be very costly when relevant objects tend to appear on lower positions of the ranked lists. To compensate for this, approximate solutions visit only a fixed number of top-ranking objects in the lists.

Systems that exploit late fusion often employ a simple monotonic aggregation function, such as sum, average or maximum, on the partial object distances. Different papers study the influence of individual functions and their parameters. The famous Threshold Algorithm for a precise aggregation of the partial results as well as a study of the theoretical behavior of the fusion process is provided in [4].

Inherent fusion In both early and late fusion, the individual modalities are treated with equivalent importance and used at the same level of query processing. For inherent fusion, one modality is chosen as primary and used to select promising data regions, typically by applying pruning techniques on the primary-modality index. The promising data are further processed with multi-modal similarity. The basic idea of this approach is the same as with postprocessing but the size of data searched with all modalities is considerably larger, thus increasing the probability of discovering more relevant objects. The inherent fusion solution allows to exploit a single-modal index structure and supports flexible weighting of modalities.

3.2.3 Selected methods

In our image-search evaluation scenario, three primary modalities are available: the text annotations, the global visual descriptors (a fixed combination of five MPEG7 features), and the local visual descriptors (SIFT). However, only the first two are suitable for the basic search, as the content-based retrieval with local descriptors is too costly for large-scale retrieval. Utilizing the MESSIF framework [1] for content-based searching and the Lucene[1] engine for text-based retrieval, we implemented all the discussed solutions, i.e. two single-modal search methods and four bi-modal solutions for basic search as depicted in Figure 2. In particular, the early fusion was implemented using a single index built for the combined text-and-visual descriptor, and the standard Threshold Algorithm was applied for the late fusion.

3.3 Postprocessing

The postprocessing phase follows the basic search and takes the query object and a set of candidates determined in the basic search as an input. The postprocessing may be applied for two different reasons: either as a part of relevance feedback loop, in which case it provides a refined query object for the next search iteration, or in the single (or final, eventually) iteration as a means of result refinement. In our study, we are only concerned with the latter alternative.

The reasons for applying the search-and-postprocess scenario instead of a more complex primary search are two-fold: 1) reduction of costs of large-scale retrieval, and 2) information mining. As we already discussed, it is too costly to employ complex evaluations of similarity over voluminous datasets, therefore only simple similarity measures are utilized in the primary search. This results in approximate searching with both false-positives and false-negatives. In the postprocessing phase, more advanced computations can be engaged on the small number of candidates, including time-demanding processing of complex distance functions or notions of similarity that are difficult to index (e.g. non-metric distance functions). While not being able to recover the false-negatives missed by the basic search, which are accepted as a toll for efficient retrieval, the postprocessing aims at eliminating the false-positives. In addition, the candidate objects produced by the basic search can serve as a source of additional information on the properties of both the query and the dataset, which in turn can be utilized to eliminate less relevant objects from the result. As the postprocessing typically produces a ranked list of the candidates, the top of which is returned to the user as the final result, this phase is also denoted as result (re)ranking.

3.3.1 Classification of approaches

In recent years, a large number of postprocessing methods have been presented in various contexts. Instead of detailing

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

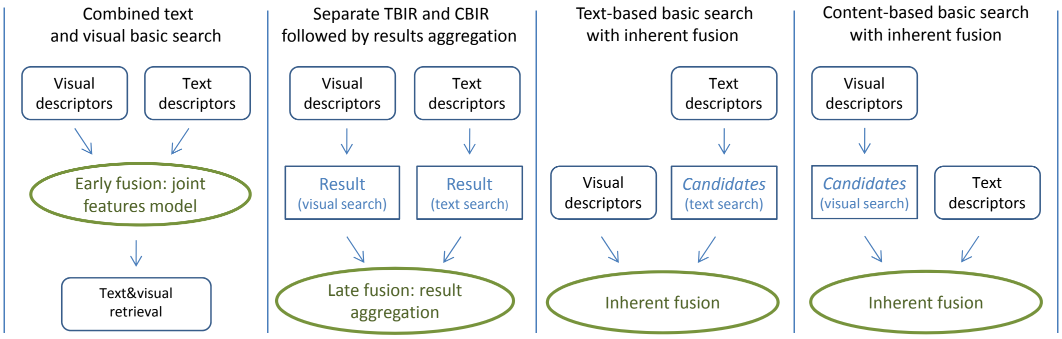
Combined text and visual basic search
Visual descriptors
Text descriptors
Early fusion: joint features model
Text&visual retrieval
Separate TBIR and CBIR followed by results aggregation
Visual descriptors
Text descriptors
Result (visual search)
Result (text search)
Late fusion: result aggregation
Text-based basic search with inherent fusion
Text descriptors
Visual descriptors
Candidates (text search)
Inherent fusion
Content-based basic search with inherent fusion
Visual descriptors
Candidates (visual search)
Text descriptors
Inherent fusion

Figure 2: Different strategies for a bi-modal basic search.

the individual approaches, we introduce the following categorization of ranking methods, taking into account the type of information exploited in the postprocessing:

Orthogonal modality ranking The candidate objects are ranked with respect to one or more modalities that were not used in the basic search. As a typical example, let us consider image search that employs primary text-based retrieval and visual reranking of the candidates.

Fusion ranking This type of postprocessing assumes two or more basic searches, results of which are taken and merged together. In contrast to the late fusion technique in the basic search phase, this solution does not require access to other data than the ranked lists of candidates. This approach is also denoted as *rank aggregation* and is typically used to combine information from heterogeneous sources.

Pseudo-RF ranking A retrieval session with relevance feedback (RF) consists of several iterations, during which users evaluate intermediate results to improve and enrich the query. In a pseudo-RF approach, the user evaluation is replaced by extracting pronounced features from the initial result set, which are expected to be important for the query. Many ranking methods work this way, exploiting properties of the initial result set. We denote the gained information as *secondary retrieval modalities*.

Interactive ranking Users are a rich source of information, both on general semantics of objects and individual preferences. In interactive search sessions, users cooperate with the retrieval system either actively (RF) or passively. In the passive mode, user behavior is monitored and the system learns to rank objects according to user's preferences. Click data or eye movements are typically studied in such cases.

Naturally, these categories are non-exclusive. In our work, we study the first three classes, discuss their contributions in different scenarios and evaluate their performance. The interactive ranking is out of the scope of our research.

3.3.2 Selected methods

For the experimental evaluation, we implemented the following postprocessing methods: 1) three representants of orthogonal modality ranking, one for each of the three modalities we consider (text, global visual descriptors, local visual descriptors); 2) a fusion ranking, which combines the results of the single-modal text-based and visual-based basic searches; and 3) two types of pseudo-RF ranking – rank by important visual descriptors determined in the candidate set, and a clustering-based ranking which favors objects near the virtual center of the candidate set.

4. EVALUATION

The eligibility of any search method is determined by two quality measures – its computational efficiency and the relevance of results. Naturally, different qualities are required by different applications. For the large-scale retrieval, efficiency and scalability are the crucial issues. Concerning the quality of search results, it is important that relevant objects are reported on the top positions of the result list; however, it is not necessary to retrieve all qualifying objects.

While the computation costs can be measured easily, the result quality evaluation is a non-trivial problem in general multimedia retrieval. Because of the complexity of the multimedia objects and their possible interpretations, we are not able to determine automatically whether an object is relevant for a given query. Therefore, user satisfaction is used to assess the relevance of objects and create the ground truth – the set of objects relevant for a query.

To test the performance of methods intended for large-scale retrieval, it is necessary to perform the evaluations over a large dataset with real-world data. As no such evaluation data was available, we decided to create a new evaluation platform as described bellow.

4.1 Data and queries

As anticipated, we evaluated all experiments over a large collection of real-world image data. In particular, we engaged the Profiset[2] data collection, which contains 20 million stock photos with rich and precise keyword annotations.

To evaluate the retrieval quality, we defined a set of 100 queries, each of which is composed of an example image and a short description. The topics comprise a selection of the most popular queries from search logs provided by a commercial partner, and several queries that are known to be either easy or difficult to process in content-based searching. Figure 3 shows a few queries from our selection.

4.2 Ground truth

The relevance of result objects was evaluated in the following way: for each query, top-30 queries were run using each of the methods, and the results were displayed to users for evaluation. Users sorted the images into three categories –

[2] http://mufin.fi.muni.cz/profiset

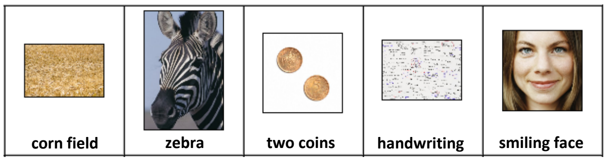

Figure 3: Query objects.

highly relevant, partially relevant, irrelevant – using a web interface. At least two users evaluated each result to compensate for subjectivity. Afterwards, the categories were transformed into percentual relevance values and averaged.

The results of all the experiments over the Profiset data and the collected the relevance assessments were made freely available to the research community as the Profiset evaluation platform [2]. The data can be used for other evaluations in future, thus sparing other research groups from the tedious labor of collecting the ground truth data and moreover, enabling fair comparison of other search methods.

4.3 Evaluation of results

To evaluate the overall performance of the individual retrieval methods, we compare both their costs and the quality of results. Concerning the search efficiency, computation time measure is the most natural choice as all the methods are run on the same hardware. As for effectiveness, we need to choose the evaluation metrics carefully as our ground truth data is not of the typical sort assumed in information retrieval – it is incomplete and with non-binary evaluations of relevance. According to the discussion in [10], we selected two quality measures – Precision at k and Normalized Discounted Cumulative Gain (NDCG) at k. These allow to evaluate the precision of the top k results retrieved by a particular method relatively to the best known results for the given query, with the latter also taking the ranking of the result objects into account. A fair comparison of the selected methods is obtained this way, even though the absolute values of the quality metrics might be different with a more complete ground truth data.

5. DISCUSSION OF RESULTS

In this section, we analyze the results of the experimental evaluation from several viewpoints. First, we consider the retrieval scenario with a multi-modal query, and study the trade-off between evaluation costs and result quality. To be able to see clearly the differences between individual approaches, we limit our view on the two modalities that can be used in all phases and combinations, i.e. the text and global visual descriptors. In the second part, we study the effects of ranking methods used with a single-modal basic search, focusing also on the benefits gained by exploiting secondary modalities that can be obtained from the initial result. We also analyze the influence of the initial result size on the performance of ranking-based methods.

5.1 Bi-modal fusion performance

The first phenomenon we study is the performance of different solutions that combine the two basic modalities used in image retrieval – the text (T) and global visual (V) descriptors. As detailed earlier, our evaluation comprises both the simple solutions based on single-modality basic search and complementary ranking, and the more expensive fusion approaches (combined indexes, Threshold Algorithm). Figures 4 and 5 show the comparison of all the bi-modal combinations in terms of response time and result precision, respectively.

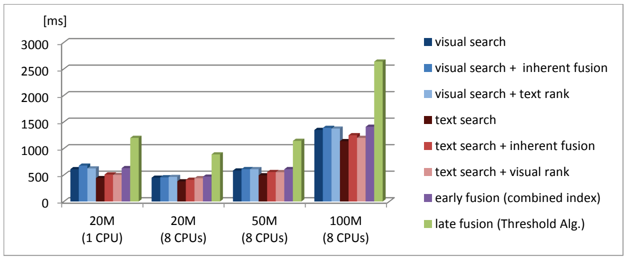

Figure 4: Response times.

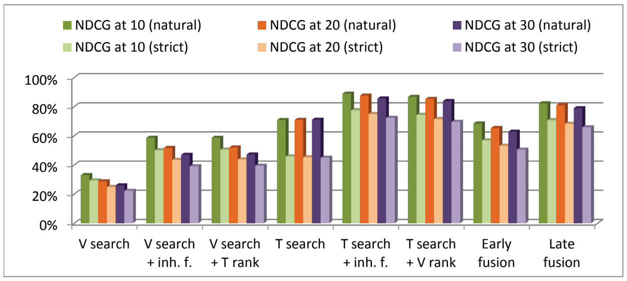

Figure 5: Performance of bi-modal search methods.

5.1.1 *Efficiency*

The performance was measured using wall-clock time needed for evaluating a single query. All experiments were run on a single machine with 8 CPU cores and 32GB RAM in two settings: 1) using only one CPU, thus obtaining the baseline costs of the index, or 2) using all CPUs. The latter approach utilized internal parallelization of the indexes, so the speedup is not proportional to the number of CPUs used, since the indexes usually have some fixed costs. We can observe the averaged response times in Figure 4. The first two entries represent the values measured in our experiments on the Profiset data. The values for the 50M and 100M datasets are estimations based on the performance of the single-modality indexes and are only provided to demonstrate the trends.

We can see that the times for a single-modal search are increased marginally when ranking or inherent fusion was used. In particular, the visual search time increased from 610 ms to 625 ms when text ranking was used and to 677 ms with inherent fusion. Similarly, we can observe a 14 % increase of the text search costs with global descriptor ranking. These differences are even lower when a parallelization is used inside the index and, in fact, inherent fusion is even more efficient than the post-processing that needs to wait for the index to supply the full result before the ranking is computed. Quite noticeable are the high costs of the Threshold Algorithm that are more than two times higher than the other variants. Despite its ability to produce precise results, the Threshold Algorithm needs to process significantly more data from both the text and visual indices. Finally, the early-fusion approach shows about 5 % higher response times than the global descriptor single-modal index.

5.1.2 *Effectiveness*

Figure 5 presents a comparison of the average relevance achieved by eight possible combinations of the text and global visual modality. It shows the NDCG metric values in two modes: *natural*, which uses non-binary relevance of objects as determined by the relevance assessments, and *strict*, where the relevance is transformed into binary values to reveal the percentage of perfect objects in the result sets.

The graph reveals that the best results are produced by the text-based basic search with inherent fusion or visual ranking – the difference between these two approaches is negligible. In comparison to the single-modal text search, the secondary modality improves the overall relevance by approximately 20 % and significantly increases the ratio of highly relevant results. Similar behavior can also be observed for visual search with text ranking or inherent fusion. The late fusion approach provides nearly as good results as the best one, but early fusion drops behind. This is caused by approximations employed by the content-based retrieval, which are also used in visual-based basic search but become more pronounced for the complex fused features.

To explain the success of text-based methods, we need to recall that the Profiset collection contains data with rich and precise annotations. In fact, the text metadata were created to maximize the findability of stock photographs. It is natural then, that text-based matches produce relevant results. We can also hypothesize that the textual (semantic) relevance is more natural to people than the visual similarity, which would also increase the score of text-based retrieval.

The suitability of text-based approaches for retrieval in collections with good text data is well known and used in commercial applications, i.e. [6]. The important phenomenon to notice is the fact that this simple, approximate approach to multi-modal retrieval outperforms even the precise Threshold Algorithm (TA), which (in theory) should provide optimal results. Probably, we could achieve better results with TA with a more finely tuned balance of the textual and visual components. However, there is no sense in doing so when the text search combined with visual ranking provides the same quality of results in much lower time. When we try to evaluate the "balanced" fusion in an approximate way – either by means of early fusion with approximate index search, or approximate TA (i.e. fusion rank, not shown in graphs) – the performance drops down.

Looking again at Figure 5, we can learn another important fact. The performance of two-phase retrieval differs significantly between the solutions exploiting text-based and visual-based primary search. Although the same two modalities are combined in both approaches, the order of their utilization has strong influence on the results. Let us denote the more successful modality as the *dominating modality*. Our experiments have shown that in case of general-purpose retrieval and image collection with high-quality text data, the textual component is the dominating modality. Automatic recognition of a dominating modality for different situations is a great challenge for future research.

5.1.3 *Limits of text-based approach*

Apart from the overall evaluation that utilized the averaged results across all queries, we also studied the result precision for individual queries. Naturally, the text-based retrieval with visual ranking does not perform optimally for all of them. So far, we have identified the following categories of queries for which the text-based approach is less suitable (or, the text component is not dominating): complex queries ("two coins"), ambiguous queries ("shells", "stamp"), and too broad queries ("bird", illustrated in Figure 6). A more detailed study of these cases will be part of our future work.

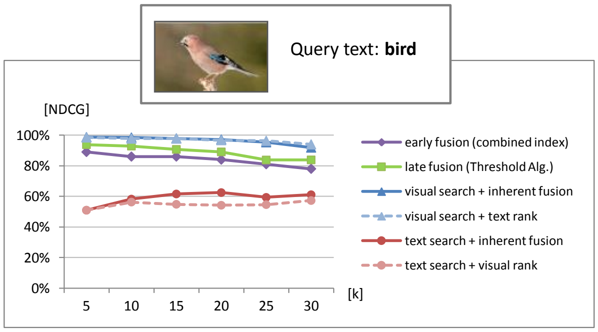

Figure 6: Limitations of text-based retrieval.

5.2 Multi-modal retrieval with ranking

Our next interest lies in discovering the suitable pairs of modalities, which provide best complements to each other. For this purpose, we study all modalities – the primary as well as secondary ones. Since we discovered in the first experiment that the multi-modal basic search does not provide interesting results, we limit our view to the ranking of results of the visual-only and text-only basic search.

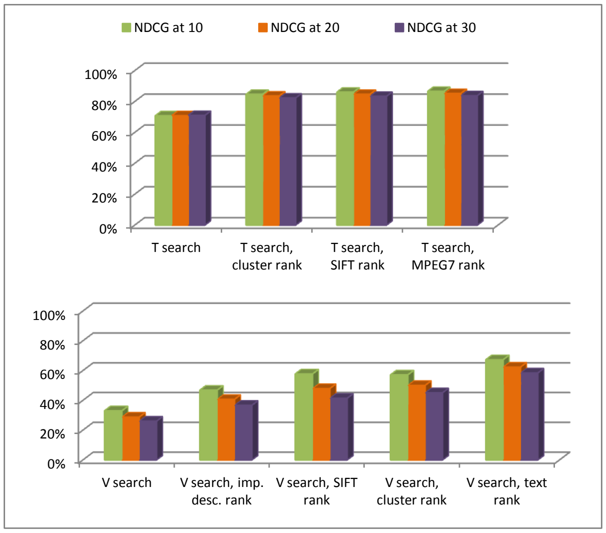

Figure 7: Text and visual search with ranking.

Figure 7 shows the performance of the combinations we examined. The first graph reveals that the most suitable complement to text search is the ranking by global visual features. This is an interesting finding, since majority of existing text-based search engines prefer local descriptor ranking. In our experience, however, local descriptors are better suited for narrow-domain search or subimage retrieval, whereas global descriptors perform well in broad domains.

The lower graph shows that the same pair of modalities performs well in an inverted order – the text-based ranking provides best improvement to visual basic search. However, we are also interested in the performance of the non-textual complementary modalities, as there are applications where visual-only query is evaluated over text-and-visual data, e.g. image annotation. In such scenarios, the local-descriptor primary modality and clustering secondary modality are the most suitable complements of the global visual descriptors. The clustering modality is especially interesting, since it allows to exploit the text component of the data objects even though text is not included in the query definition.

5.2.1 *Influence of initial result size*

An obvious question related to the two-phase search model concerns the choice of the initial result size. This needs to balance two factors: the quality of the final results and the costs of the postprocessing phase.

In our experiments, we have evaluated all the postprocessing methods with three different settings of the initial results size: 100, 500 and 2,000 objects. From the efficiency point of view, the difference between the costs was insignificant. However, distinct trends could be observed concerning the quality of results. Interestingly, the trends differed for the ranking of results obtained by text-based initial search, and visual-based primary search. For the solutions that exploit text as the primary modality, the quality of results continues to grow with the size of the initial result, even though the improvements are less pronounced for the larger numbers. We could also observe this trend in Figure 5, where the text search with inherent fusion, which ranks up to 15,000 objects, outperformed the text search with mere ranking. However, the ranking of visual-based initial results does not follow the same pattern. There is an increase of result quality between initial results sized 100 and 500, but the ranking applied on 2,000 objects performs worse than with 500. Thus in this case, the inherent fusion is counter-productive.

Naturally, this phenomenon calls for further experiments, which would identify the factors that determine the most suitable sizes of the initial result. Our experience suggests that the different behavior of text-based and visual-based initial results could be related to the intrinsic dimensionality of the respective search spaces. We intend to study these relationships more deeply in future research.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have established a classification of approaches to multi-modal image retrieval, and analyzed the results of our first experiments aimed at deciding the suitability of the available solutions for different use-cases. For the purpose of multi-modal image retrieval, we tested some of the hypotheses that were generally accepted by the research community but so far not evaluated in a large-scale environment. We confirmed that the two-phase retrieval with the text-based initial search is the most eligible method for image retrieval in case of simple queries and target dataset with high-quality text data, which is an accord with analyses performed within the ImageCLEF evaluation campaigns [12]. However, we were able to identify several types of queries that need special attention and further study. We also provide insights into the mutual relationships of different modalities and the performance of the most frequent combinations, with a special attention to solutions that can be utilized in automatic image annotation.

Although we have been able to identify some of the mechanics of the multi-modal searching, a lot of issues still remains to be studied in the future. Some of the topics were mentioned in the discussion of results, in particular the identification of dominating modality, or the analysis of the relationship between the intrinsic dimensionality of the search space and the performance of the ranking methods. We also plan to evaluate the same set of experiments over a dataset with low-quality text data and study the performance of the chosen methods in a different situation.

Acknowledgments

This work has been partially supported by Brno Ph.D. Talent Financial Aid and by the national research projects GD 102/09/H042 and GAP 103/10/0886.

7. REFERENCES

[1] M. Batko, D. Novak, and P. Zezula. MESSIF: Metric similarity search implementation framework. In *First International DELOS Conference*, volume 4877 of *LNCS*, pages 1–10. Springer, 2007.

[2] P. Budikova, M. Batko, and P. Zezula. Evaluation platform for content-based image retrieval systems. In *TPDL 2011*, pages 1–12, 2011.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40:5:1–5:60, May 2008.

[4] R. Fagin. Combining fuzzy information: an overview. *SIGMOD Rec.*, 31:109–118, June 2002.

[5] R. Jain and P. Sinha. Content without context is meaningless. In *ACM Multimedia 2010*, pages 1259–1268, New York, NY, USA, 2010. ACM.

[6] Y. Jing and S. Baluja. VisualRank: Applying PageRank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1877–1890, 2008.

[7] J. Kludas, E. Bruno, and S. Marchand-Maillet. Information fusion in multimedia information retrieval. In *AMR 2007*, volume 4918 of *LNCS*, pages 147–159. Springer Berlin / Heidelberg, 2008.

[8] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2:1–19, February 2006.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[10] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, 2008.

[11] MPEG-7. Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002, 2002.

[12] H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, volume 32 of *The Information Retrieval Series*. Springer, 2010.

[13] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search – The Metric Space Approach*, volume 32. Springer, 2006.

[14] X. Zhou, A. Depeursinge, and H. Müller. Information fusion for combining visual and textual image retrieval. In *20th International Conference on Pattern Recognition*, pages 1590 –1593, aug. 2010.