Large-scale multi-modal image search: theory and practice

Petra Budikova

Michal Batko

David Novak

Pavel Zezula

Abstract

One of the major challenges of modern data processing is to provide an efficient multi-modal retrieval in large collections of complex data objects. In this paper, we study different late fusion solutions for image searching and analyze their performance in terms of effectiveness, efficiency, and scalability. We introduce a classification of existing approaches and present a novel technique of inherent fusion that combines the efficiency of fast indexed retrieval with the effectiveness of ranking methods. The performance of all these methods is evaluated in extensive experiments with user participation.

Introduction

With the rapid growth of both volume and diversity of digital data, traditional attribute-based or text-based retrieval does not always provide satisfactory results and more sophisticated approaches to data management are needed to meet users’ information needs. This applies for many complex data domains and in particular for multimedia content, which is becoming very popular in many applications. For image data alone, the use cases range from scientific data management to entertainment, security, and surveillance. In all these tasks, efficient and effective retrieval methods are needed that are able to take into account different aspects of the image content and to adjust the searching according to particular user’s preferences.

The first approaches to image searching followed the established lines of attribute- and text-based 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 or not sufficient in terms of quality and descriptiveness. 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 (MPEG-7, 2002)), local image features (e.g. SIFT (Lowe, 2004)), face descriptors, etc. However, the usefulness of such solutions is limited by the *semantic gap* problem, which refers to the disproportion between the object similarity on the level of selected features and the human-perceived relationship between objects, which takes semantics into account (Datta et al., 2008).

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 a 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) 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 irrelevant ones. Therefore, latest data management techniques focus on a *multi-modal retrieval* that combines multiple orthogonal views on objects (Datta et al., 2008; Jain and Sinha, 2010; Kherfi et al., 2004).

Already, a number of solutions for multi-modal image retrieval exist and they are rapidly developing. Each such solution is suited for a certain type of applications that has its specific characteristics and requirements. In this study, we focus on a general-purpose searching in very large collections, which could be applied e.g. in a web search engine. Even for this specific subproblem, many diverse research directions exist that are pursued by different research communities with different backgrounds. To mention at least two examples from the opposite ends of the spectrum, there is the practical approximate Google image search based on text retrieval with content-based *re-ranking* (Jing and Baluja, 2008), and the precise theoretical Threshold Algorithm for *fusion* of multiple single-modal search results (Fagin, 2002).

The existence of diverse research directions in image retrieval is very natural. However, to achieve real improvements and mature solutions, it is also necessary to have a cooperation and comparison between individual approaches. Unfortunately, this is rather scarce in this area due to the lack of commonly accepted benchmarking platforms (Lew et al., 2006). The research groups tend to work with their own special datasets, application settings etc., making the presented results virtually incomparable.

Objectives

In this paper, we focus on the issues related to providing efficient and effective solutions for web-scale image retrieval. In this context, the most important qualities of the retrieval process are its speed, scalability, and flexibility. We study the behavior of various multi-modal retrieval methods with respect to these criteria and propose a novel approach that combines the advantages of several previous solutions. Then, we perform an extensive experimental evaluation of the fusion techniques over a large-scale, real-world image dataset. In the evaluation, we employ two popular modalities of image retrieval – global image content descriptors, and text annotations (Datta et al., 2008).

The main contributions of this paper are the following:

* Systematic categorization of approaches to the multi-modal retrieval and a survey of existing techniques: We formalize the multi-modal retrieval problem and present a categorization of possible approaches. Then, we discuss the theoretical applicability of selected state-of-the-art solutions.
* Specification of a novel fusion technique: We introduce the *inherent fusion* technique, which provides efficient and flexible multi-modal retrieval with high precision. We outline the general principles of this method and discuss the practical implementation aspects.
* Effectiveness and efficiency comparison of various approaches to multi-modal retrieval: Using a novel evaluation platform, we study the performance of various query processing methods. In particular, we analyze the trade-off between search costs and precision for different fusion scenarios, and the performance of individual methods for different types of queries.

The paper is based on our previous work (Budikova et al., 2012), which is extended by a detailed description of the inherent fusion method, and a more thorough analysis of experimental results.

Flexible multi-modal retrieval

Multi-modal approach to multimedia data retrieval is widely accepted as a promising way of overcoming limitations of mono-modal approaches and obtaining semantically relevant results within acceptable retrieval costs. In recent years, many studies of possible multi-modal approaches have been performed and numerous fusion methods have been proposed. A thorough discussion of all aspects of modality fusion is out of scope of this paper, therefore we refer readers to survey studies (Atrey et al., 2010; Bozzon and Fraternali, 2010; Depeursinge and Müller, 2010) to get a more complex view of the problem. In this work, we leave aside the problems of suitable selection and preprocessing of modalities, and focus on the actual process of fusion. After a brief introduction of the basic principles, we focus on late fusion methods that allow flexible similarity searching.

Multi-modal search model

For further discussions, let us first clarify the concept of modality. Informally, a modality can be understood as a point of view, which transforms complex objects from a given dataset  to representations more suitable for searching. Formally, we define modality in the following way:

**Definition**

A modality  is represented by a projection function ,  being a domain of modality , and a distance function . The projection function transforms an object into a feature descriptor , while the function  evaluates the distance between two descriptors, i.e. the dissimilarity of two objects as seen in the view of the modality .

Multi-modal retrieval paradigm assumes the existence of multiple modalities , which can be utilized to manage data from . In some moment of the data processing, these modalities are combined – *fused* – to provide more complex representations of objects from  and to evaluate their similarity on a higher semantic level. Let  be a multi-modal search engine that recognizes modalities .  provides the following tools to manage data objects from  : a set of supported multi-modal projection functions , and a set of available multi-modal distance functions . The functions from  and  are utilized for data indexing and retrieval.

In this study, we focus on  *nearest neighbors* (*NN*) queries, which are most common in multimedia retrieval. A query  over  is defined by a multi-modal query object  and a distance function . The query distance  needs to be taken from the set  of supported distance functions. The semantics of *kNN*(*Q*) is the following: given the query object , the system returns the  objects from  that are most similar to  with respect to the query dissimilarity measure .

Let now suppose a multi-modal search system that organizes the dataset . The system can be designed in two basic ways (or, eventually, a combination of both):

* *Early fusion* approaches combine modalities  prior to data indexing. The fusion is evaluated off-line, therefore extensive analysis of relationships between the individual modalities can be performed to achieve better understanding of data semantics. Early fusion produces a new complex descriptor with an associated overall distance measure, which is employed to index and search the data.
* *Late fusion*, on the other hand, fuses modalities during query evaluation. The data is organized in one or several index structures that correspond to individual modalities. During the actual search, candidate objects are identified in these separate indexes and their similarity is then re-evaluated (fused) using the query distance measure .

The objective of this work is to study large-scale image retrieval over broad domains. In this context, one of the most important characteristics of a successful search system is its flexibility – users should be able to adjust the similarity measure  to their individual needs in a given situation. Early fusion techniques typically allow to search data only by the indexing distance function and therefore are not suitable for such tasks. Late fusion in principle does not pose any restrictions on  and is thus better suited for adaptive searching. Therefore, we limit our discussions to the late fusion approach in the following analysis.

Classification of late fusion methods

Late fusion methods can be further classified with respect to two aspects that influence both the quality of results and the query processing efficiency: 1) the approach to modality integration, and 2) the number of objects accessed in the fusion phase. We explore each of these dimensions in more detail.

Integration of modalities

Late fusion is sometimes also denoted as *decision-level fusion* in contrast to *feature-level* early fusion. Indeed, the late fusion operates over results – decisions – of earlier query processing phases. These decisions take form of mono-modal object-query distances or candidate sets selected in mono-modal index structures. There are two principal ways in which late fusion may treat the modalities, which we denote as *symmetric* and *asymmetric* solutions:

* In the symmetric approach, all modalities  are exploited for parallel retrieval of candidate sets , which are then merged into a single candidate set and re-evaluated with respect to the query distance . All modalities are thus exploited at the same time and with the same level of importance – the  function is the only factor that influences the semantics of the final result.
* In asymmetric fusion, the situation is different. A subset of the available modalities is selected as *primary* (denoted as ) and used to retrieve the candidate set , which is further processed. The remaining (secondary) modalities, denoted as , are less influential than the primary ones – an object that does not rank high with the primary modalities will not appear in , even if it would have been considered perfect by some of the secondary modalities.

The reason for the utilization of asymmetric solutions may be threefold: the primary modalities may really be more vital for a given use-case scenario; the asymmetric solution may be chosen because of efficiency issues; or some of the modalities may not be available at the beginning of the query evaluation. We shall explore the latter two reasons in more detail in the following sections.

Fusion scenarios

The second classification dimension we would like to discuss concerns the number of objects accessed during the fusion processing. This issue is closely related to the precision of the retrieval result with respect to the similarity measure . We find it helpful to distinguish two phases of the query evaluation that may accommodate the fusion:

* *Basic search phase* is the core part of query evaluation that accesses index structures and identifies candidate objects. Processes evaluated in the basic search phase have access to all objects from the dataset . For fusion, this means that as many objects as needed can be accessed to provide best possible results with respect to .
* *Postprocessing phase* follows after the basic search and evaluates additional computations over a set of candidate objects , provided by the basic search. If the modalities are fused in this phase, the precision of the final result is influenced by the selection of  – objects that have been discarded in the (mono-modal) basic search cannot appear in the result. The precision of fusion evaluated in the postprocessing phase is thus limited by the performance of modalities that provide the candidate set. On the other hand,  is typically much smaller than  (usually several hundreds or thousands of objects), therefore the fusion costs are relatively low and independent of the dataset size.

Obviously, the quality of fusion evaluated in these two phases may differ significantly in both the processing costs and results precision. Moreover, the postprocessing fusion may exploit the properties of candidate objects in  to extract some additional information, which can be utilized in various (pseudo)-relevance feedback strategies to improve the relevance of the answer set.

State-of-the-art in fusion techniques

In the following sections, we study state-of-the-art search methods belonging to different classes defined by the two classification dimensions we introduced. We describe some of the most influential solutions and discuss the advantages and limitations of different fusion designs. We mainly focus on large-scale image retrieval and the two modalities most frequently used in this context – visual content descriptors, and textual image annotations.

Symmetric fusion in basic search phase

Precise symmetric late fusion is best represented by the Threshold Algorithm, which was introduced by Ronald Fagin in 2002 (Fagin, 2002). Partial results provided by individual modalities are aggregated in the basic search phase, accessing as many objects as necessary to guarantee precise fusion results. The algorithm is iterative and works as follows: Let  be the input modalities and  be ordered lists of objects from  defined by individual modalities. Objects in  are ordered by increasing distance from  as measured by . In each iteration, the Threshold Algorithm takes the top unvisited object from each sorted list, adds it to candidate set , and evaluates its overall distance  from . Then, a threshold condition is verified which decides whether the top  objects in represent the best possible result, or another iteration needs to be executed.

The Threshold Algorithm represents a theoretically precise, clear solution that is applicable in many situations. It allows to combine results of independent search systems, which can be queried in parallel for the sorted lists. The aggregation function needs to be monotone, but this property holds for most aggregations that are used in real search systems. Unfortunately, there are no reasonable limitations of the fusion processing costs. In the worst case, it is possible that the algorithm will need to visit all objects in the database to be sure that the optimal solution was found, which is not acceptable in large-scale applications.

Asymmetric fusion in basic search phase

Asymmetric fusion in the basic search phase assumes that a significant amount of objects from  is evaluated with respect to all modalities, but only a subset of available modalities is used to organize the dataset. Such approach requires either specialized index structures, or specialized retrieval algorithms that are able to operate on top of a standard mono-modal index.

To the best of our knowledge, solutions of this type have not yet been used in image retrieval but are studied in other domains, e.g. spatio-textual similarity search. In particular, the IR-tree (Cong et al., 2009) extends the standard R-tree spatial index to store both spatial and text information about points of interest. Non-leaf nodes of the IR-tree contain summarized information about text data in respective subtrees, which allows a search algorithm to prune the search space efficiently with respect to both textual and spatial modalities. Similar to the Threshold Algorithm, the aggregation function needs to be monotone.

The IR-tree enables precise and efficient asymmetric fusion, but is designed to support only the two specific data modalities. Providing a general asymmetric basic-search solution remains a challenge, which will be addressed in the next section.

Symmetric fusion in postprocessing phase

Symmetric postprocessing fusion is actually an approximation of the Threshold Algorithm that accesses only a fixed number of objects from each sorted list  provided by modality . The reasons for exploiting this approximation may be threefold: 1) the precise modality fusion evaluated by the Threshold Algorithm is too expensive, 2) the requested aggregation function is not monotonic, or 3) the mono-modal search systems that provide input for the fusion phase do not offer full sorted lists of objects from .

Experiments over real-world data, discussed e.g. in (Batko et al., 2008), reveal that the Threshold Algorithm requires many iterations to reach the optimum result, but most of these iterations bring little improvement of the result quality. Therefore, the authors propose to fuse only a limited number of objects and provide users with an estimate of the result quality. Many other solutions take the advantages of postprocessing fusion for granted and focus on refining the aggregation rules. In (Liu et al., 2009), the CrowdReranking algorithm is presented, which combines results of multiple text-based web search engines to increase the relevance of text retrieval. The aggregation works on a voting principle known from classification algorithms. Application of fuzzy inference rules for scores fusion is proposed in (Chatzichristofis et al., 2012). Symmetric postprocessing fusion is also frequently utilized in solutions of the ImageCLEF tasks (Depeursinge and Müller, 2010), which typically exploit linear combinations of modalities.

Easy applicability of the fusion phase on top of existing search systems and low additional costs are two obvious advantages of all postprocessing solutions. On the negative side, these techniques produce approximate results with no quality guarantees. Under certain conditions, the precision loss may be insignificant – a certain level of imprecision is inherently contained in content-based data management and some false dismissals of relevant objects are acceptable, especially in large data collections. However, finding the optimum balance between retrieval costs and results quality is still an open problem.

Asymmetric fusion in postprocessing phase

Asymmetric postprocessing fusion represents the approximate alternative to precise asymmetric solutions. One or several primary modalities are exploited to provide the set of candidates , which is then re-evaluated with respect to additional (or all) modalities. This approach is also denoted as *result re-ranking*, since the ordering of candidates in  is typically updated in the postprocessing and the top-ranking objects are reported as the final result. Apart from the obvious option of re-ranking by a modality orthogonal to the primary ones, the asymmetric postprocessing also provides space for relevance feedback (RF) and pseudo-RF processing.

Ranking by orthogonal modality is well known from commercial image search systems Google (Jing and Baluja, 2008) or Bing (Wang et al., 2009), both of which exploit traditional text retrieval to obtain the candidate objects and then reorder the results with respect to visual similarity. A random walk over a visual similarity graph is exploited to determine the final ranking of results. A complementary approach that exploits visual features as the primary modality is presented in (Budikova et al., 2011), which proposes several techniques for textual ranking of  and discusses pseudo-RF ranking techniques that exploit information learned from objects in . Pseudo-RF postprocessing is also utilized in (Chen et al., 2010) to adjust the weights of multiple visual features that are fused. Solutions presented by Zitouni et al. (2008), Hörster et al. (2009) or Mironica et al. (2012) apply various types of clustering, giving higher ranks to large clusters or clusters which have a centroid nearest to the query object. Finally, Jegou et al. (2010) propose to use the reverse-kNN query and increase the rank of objects that have the query among their nearest neighbors.

Re-ranking solutions are popular among contemporary multimedia retrieval systems as they can be implemented directly on top of an existing mono-modal retrieval system, e.g. a text-based search engine. The query processing can be very cheap if efficient index structures are available for the primary modalities. The RF and pseudo-RF ranking strategies also provide strong tools for overcoming the semantic gap problem. On the other hand, the performance of asymmetric fusion strategies strongly depends on the quality of candidate set provided by primary modalities. Therefore, the applicability of such solutions is limited to datasets where the primary modalities are available in sufficient quality.

Inherent fusion

The key factor of the asymmetric postprocessing fusion is the fact that the candidate set  that is passed to the ranking phase has a preset size. If this size is too large, the search performed on the primary modality can be very costly. Moreover, the candidate set has to be fully enumerated, which may require additional memory and communication costs. On the other hand, if the size is too small, the overall result will be strongly affected by the primary modality, since the objects that would be highly ranked by the secondary modality are unlikely to appear within the candidates.

We believe that it is possible to significantly improve the performance of asymmetric solutions, if we more thoroughly exploit all information available during the query evaluation. Let us first have a closer look at the processing in the basic search phase. In general, indexing techniques typically partition the dataset into a number of, not necessarily disjoint, “data chunks” (intervals, areas, clusters, posting lists, etc.); let us denote these partitions , where . During the evaluation of a query, the index typically prunes some of these partitions and accesses the potentially relevant data in the rest of them to select the answer set. The number of objects accessed in this way is significantly (orders of magnitude) larger than the typical number of candidates .

Following this observation, let us propose a technique of *inherent fusion* that utilizes all objects visited by the index also for ranking by secondary modalities. Similarly as in the asymmetric fusion techniques, we propose to index the data using selected primary modalities  but store the full data objects, so that all primary and secondary modalities ( and ) are held by the index. At query time, the index processes all objects from all non-pruned partitions; let us denote these objects as “super-candidate” set defined as , where  is a data partition that cannot be pruned for the query object . Instead of a standard accumulation of the best-seen objects with respect to the primary modality , each object from  is evaluated by the multi-modal similarity function  which takes into account both the primary and secondary modalities.

The difference between the standard asymmetric postprocessing fusion and inherent fusion is schematically shown by Figure 1. In case of the postprocessing (left schema), the index on  identifies relevant partitions  and ranks the data from these partitions by  to create the candidate set for further processing by the query distance. As described above, the inherent fusion (Figure 1, right) ranks the objects directly by  as they are accessed in the partition. In comparison with the postprocessing fusion, the volume of data searched with all modalities () is considerably larger, increasing the probability of discovering more relevant objects. Importantly, this processing is far less costly than the asymmetric postprocessing fusion on a candidate set of the same volume because that would typically require processing of even larger  within the  index. Obviously, it is not guaranteed that  contains the  best objects with respect to ; on the other hand, this approximation allows us to keep the processing costs low.

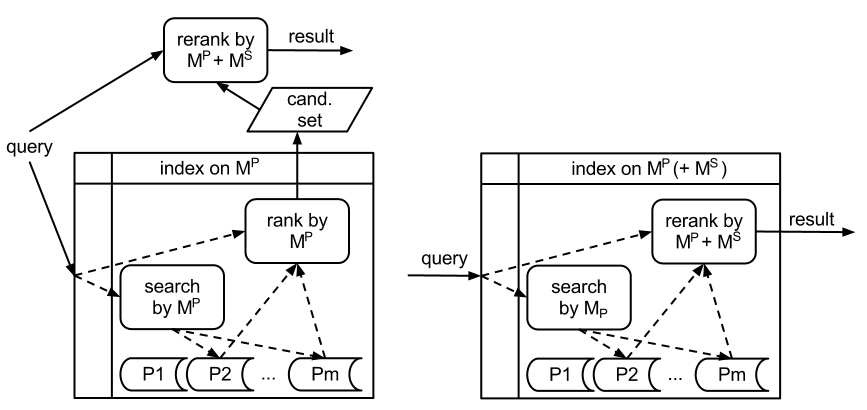


Figure 1: Schema of asymmetric postprocessing fusion (left) and inherent fusion (right).

The inherent fusion can be implemented relatively easily within most of the standard indexing techniques. First, the index needs to store the data for the secondary modalities  so that the ranking can be applied, but storing additional (raw) data is usually supported. Second, the query evaluation procedure needs to be modified, so that additional computation can be added to the part where the resulting set is accumulated during the processing. This might be possible to register via hooks (callback methods), if the implementation allows it, or the code must be modified accordingly. Finally, the system must be modified so that the query primary and secondary modalities are split and passed to the original index partition traversal or the modified result set accumulator respectively. In our implementations, we utilize the MESSIF library (Batko et al., 2007), which contains all the necessary support, so any index structure implemented on top of MESSIF can transparently take advantage of the inherent fusion.

The inherent fusion is a straightforward extension of the re-ranking paradigm, however the advantages obtained by this solution are considerable. We review them with respect to the standard quality measures of retrieval methods:

* flexibility: similarly to the standard re-ranking, there are no requirements on the way in which modalities  and  are combined at query time (for instance, arbitrary weighting) and there are no limitations on the indexability of the additional modalities ();
* effectiveness: relatively large set of objects can be probed with all modalities, which is likely to improve the quality of the results;
* efficiency: the whole evaluation is done within the index without explicit enumeration of the whole candidate set  and without any data replication, which allows us to keep the processing costs low;
* scalability: this approach allows efficient exploitation of distributed indexing techniques; scalability of the index structure is thus straightforwardly exploited to guarantee also the scalability of the inherent fusion.

The only disadvantage of the inherent fusion solution that we are aware of is the fact that the data stored in the index must contain also the secondary modalities, so the index requires more storage space. On the other hand, the secondary-modality data needs to be stored in any case for the postprocessing phase, as an online extraction of the respective descriptors would be too costly. The difference is thus only in the implementation of the storage facilities.

Evaluation methodology

The eligibility of any search method is strongly influenced by two natural 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.

In this study, we are interested in a comparative evaluation of performance of late fusion methods presented in the previous sections. 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 measured 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. In our experiments, we use the Profiset platform (Budikova et al., 2011) that was recently created to support large-scale evaluations of image retrieval systems. In this section, we discuss the selection and implementation of methods we compared, and describe the evaluation testbed.

Selected methods

In the introduction, we pointed out that providing comparisons between different image retrieval strategies is one of the important issues in contemporary research in the field of multimedia retrieval. To address this problem, we perform an extensive evaluation of performance of late fusion methods. Clearly, it is not possible to implement and compare all solutions that have ever been proposed. To keep the task feasible, we only consider basic modalities and the most fundamental search strategies. We believe that such evaluation is much needed and will lay foundations for future more advanced analyses.

In particular, we limit our study to the two modalities most frequently found in image retrieval, i.e. the text similarity of keyword image descriptions and the global visual similarity of image content. The text similarity is expressed by the cosine distance and standard *tf-idf* weighting scheme (Baeza-Yates and Ribeiro-Neto, 2011), whereas the visual similarity is evaluated by a static combination of selected MPEG-7 descriptors (Batko et al., 2010). As for the search methods, we leave aside pseudo-RF techniques and focus on the performance of basic fusion scenarios. We are particularly interested in the comparison of precise and approximate solutions with different approaches to the integration of modalities, and the differences between text-based and visual-based solutions in case of asymmetric fusion scenarios. With respect to these objectives, we selected the following methods for the experimental comparison:

* baseline solutions: text-based retrieval, content-based retrieval;
* symmetric basic-search fusion: precise Threshold Algorithm;
* symmetric postprocessing fusion: approximate Threshold Algorithm with fixed sizes of ;
* asymmetric basic-search fusion: text-based retrieval with inherent fusion, content-based retrieval with inherent fusion;
* asymmetric postprocessing fusion: text retrieval with multi-modal re-ranking, content-based retrieval with multi-modal re-ranking.

To guarantee a fair comparison of the selected techniques, all of them were implemented in a uniform environment of the MESSIF framework for similarity searching (Batko et al., 2007). In particular, the M-index structure (Novak et al., 2011) was employed to support the content-based retrieval, and the Lucene engine (McCandless et al., 2010) was utilized for the text-based searching. For approximate solutions, several settings of the sizes of  and  were tested to discover the dependence of result characteristics on these parameters.

Data and queries

As anticipated, we evaluated all experiments over a large collection of real-world image data. In particular, we engaged the Profiset (Budikova et al., 2011) data collection, which contains 20 million stock photos with rich and precise keyword annotations. This high-quality data collection was intentionally selected so that we could analyze the performance of fusion methods in optimal conditions. In future, we also plan to evaluate the same set of experiments over some more erroneous dataset.

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 2 shows a few queries from our selection.

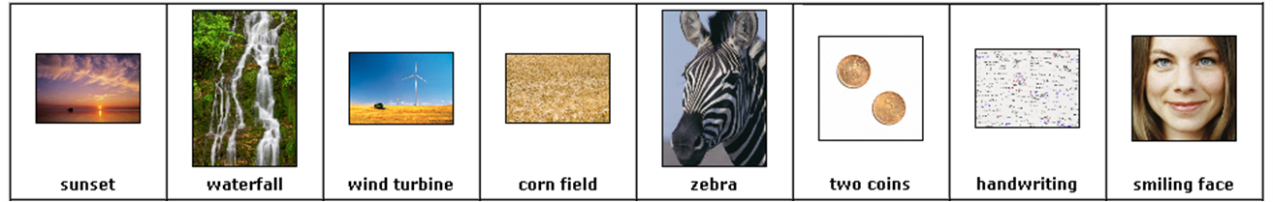


Figure 2: Query objects.

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 parameter settings, and the results were displayed to users for evaluation. Users sorted the images into three categories – highly relevant, partially relevant, and irrelevant – using a web interface. At least two users evaluated each result to compensate for subjectivity. Afterwards, the categories were transformed into percentage of relevance and averaged. This way, we obtain a *partial ground truth* for the given query objects, which can be utilized for relevance evaluations. The partial ground truth guarantees a fair comparison of the selected methods, even though the absolute values of the quality metrics might be different with a more complete ground truth data.

The results of all experiments over the Profiset data and the collected relevance assessments were made freely available to the research community as the Profiset evaluation platform (Budikova et al., 2011). 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.

Discussion of results

In the above described experiments and during the ground truth collection process, we acquired large amounts of real-use data that concern different aspects of query processing and result quality. This data allows us to perform extensive analyses of retrieval behavior of various approaches. In this work, we particularly focus on the following three subproblems:

1. How do different text-and-visual late fusion methods perform (in terms of both effectiveness and efficiency) in a large-scale real-world environment?
2. What improvements does the inherent fusion technique achieve when applied over real data?
3. What is the most suitable multi-modal search solution for a large-scale image database with high quality text descriptions?

To answer all these questions, we study three aspects of the retrieval process: the objective result quality, as measured by the distance function; the subjective result quality as perceived by users; and the query processing costs measured by wall-clock time. To the best of our knowledge, such large and comprehensive study of multi-modal retrieval with human-evaluated relevance assessments has not been done yet.

Aggregation function tuning

As discussed earlier, we currently limit our study to two modalities, which express textual and visual similarity of images. To allow multi-modal query processing, we further need to specify how these modalities should be combined. In this section, we briefly comment on the choice and tuning of the aggregation function that was applied in the experiments to facilitate the actual fusion.

Even though late fusion methods principally allow users to define (or at least, adjust) the aggregation function, in our experiments the aggregation needed to be fixed to allow a fair comparison of examined methods. We decided to employ a simple linear combination of the mono-modal distances, which is a straightforward solution that has been successfully applied in many other fusion scenarios (Depeursinge and Müller, 2010). Both visual- and text-induced distances were first normalized, and the linear fusion was tested with several weight settings. Interestingly, a balanced combination of modalities achieved best results for both symmetric and asymmetric fusion solutions, even though asymmetric approaches typically give more weight to secondary modalities in the postprocessing fusion phase (we shall discuss this phenomenon in more detail later). Therefore, the balanced combination is considered in all following comparisons.

Comparative analysis of late fusion techniques

In the first set of evaluations, we compare the performance of standard late fusion methods. Precise late fusion is represented by the Threshold Algorithm (TA), within approximate solutions we study three methods – approximate TA, re-ranking solutions based on visual (V) modality, and re-ranking based on text (T) initial search. Apart from the performance-costs trade-off, we are interested in the effects of filtering by primary modality employed in the asymmetric fusion techniques.

Distance-based result quality

A distance-based evaluation of result quality is an objective method of effectiveness assessment that compares the result of a given approximate search technique  to a precise retrieval result . From several commonly used distance-based quality measures (Zezula et al., 2006), we chose the *relative error on distance at k*, which compares the distances of objects at *k*-th position () in approximate and precise results: .

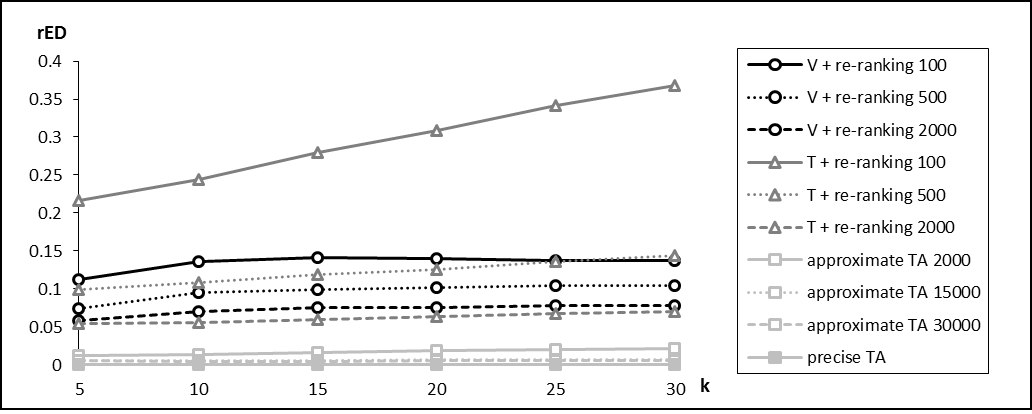


Figure 3: Standard late fusion: Average distances at a given rank.

The results provided by the *rED*(*k*) measure are depicted in Figure 3. Naturally, the precise TA shows zero error, since from the distance point of view there can be no better results. For the approximate fusion methods, the number in the method label represents the size of initial result set  which enters the postprocessing (fusion) phase. We can notice that text-based initial search followed by a re-ranking of small  has by far the worst results, which suggests that the top results of text search are not much relevant from the visual perspective. With the increasing size of , the retrieval accuracy gradually improves for all approximate techniques. For the comparable size of the candidate set  the approximate TA outperforms the asymmetric techniques, since it considers top-ranking objects from both modalities and not just the primary. However, we can observe that increasing the candidate set size beyond 2000 improves the quality only marginally.

User-perceived quality

The second quality evaluation takes into account the user-decided relevance of results. In an ideal case of a perfect distance function that precisely captures user’s information need, the user-perceived quality would copy the distance-based evaluation. In reality, however, these two perspectives may significantly differ. The *normalized discounted cumulative gain at k* (*NDCG(k)*) measure (Järvelin and Kekäläinen, 2002) considers the relevance scores of objects on positions 1 to *k*, giving more weight to higher ranking results. The value *NDCG*(*k*)=100 % corresponds with the best possible result with respect to the available partial ground truth. We apply this measure in two modes: in the *natural* mode, objects that were marked as *acceptable* during user relevance assessments are considered to have non-zero relevance, whereas in the *strict* mode, only objects that were marked as *perfect* are deemed relevant. The strict measure thus represents a more demanding user.

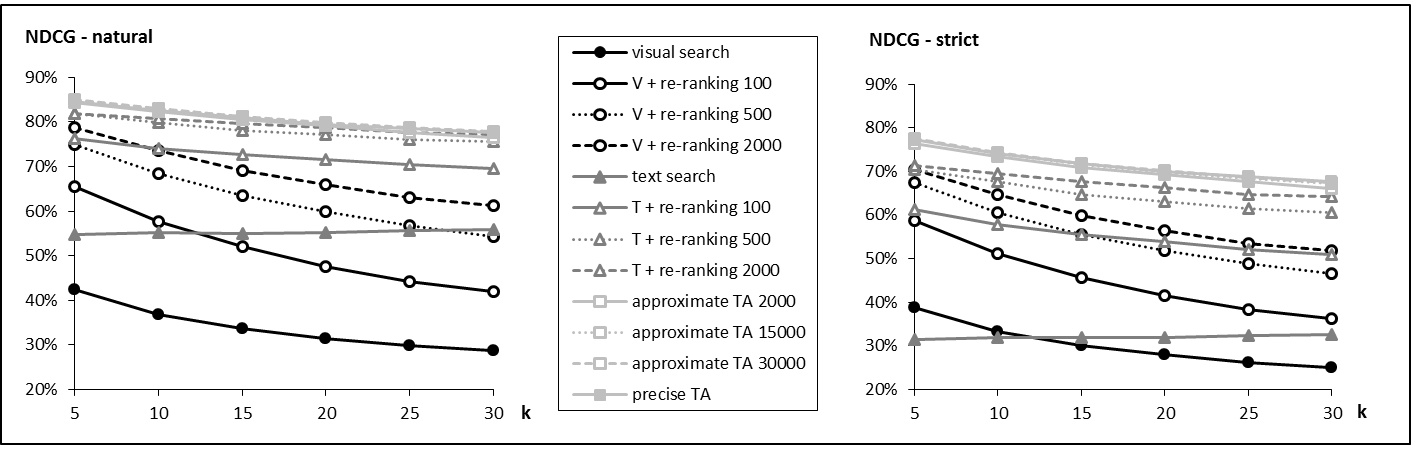


Figure 4: Standard late fusion: Average NDCG at a given rank.

In Figure 4 we can observe some differences from the distance-based evaluation. The precise TA still provides the most relevant results and is closely followed by symmetric approximations, but there is a significant difference between the user-perceived performance of text- and visual-based re-ranking methods. We were able to identify two factors that increase the success of text-based approaches: 1) users tend to prefer the semantic relevance, which is typically contained in the text descriptor, over the visual similarity; 2) the text modality is more selective – there is a distinct diversification of relevant and irrelevant objects, and the irrelevant cannot enter the postprocessing phase, whereas for visual modality there is no such clear cut.

Comparing the single-modality retrieval baselines, i.e. the visual search and the text search, we can see that practically all combined results achieved significantly higher quality. It is also of interest that the relevance of results continues to increase with the growing size of . In several previous studies including (Budikova et al., 2012), different trends were observed – the performance of asymmetric solutions began to decrease after some optimal (relatively low) value of  was exceeded. However, these solutions only applied the secondary modality to determine the ranking in the postprocessing phase, whereas in our experiments, we utilized the aggregated distance. Ranking by aggregated distance prevents objects relevant only in the secondary modality from getting to the result set, which is a desirable behavior according to our results.

Efficiency

The efficiency of the various methods has been measured using the wall-clock time needed for evaluating a single query. Since we have all the methods implemented using on the same framework and all experiments were run on a single machine with 8 CPU cores and 32GB RAM, the time is the most fair comparison method as it inherently incorporated all the evaluation aspects. In order to obtain the baseline costs of each method, we have first run the experiment using only one CPU. In the other set of experiments we have used all 8 CPUs and thus allowed the indexes to utilize their internal parallelization.

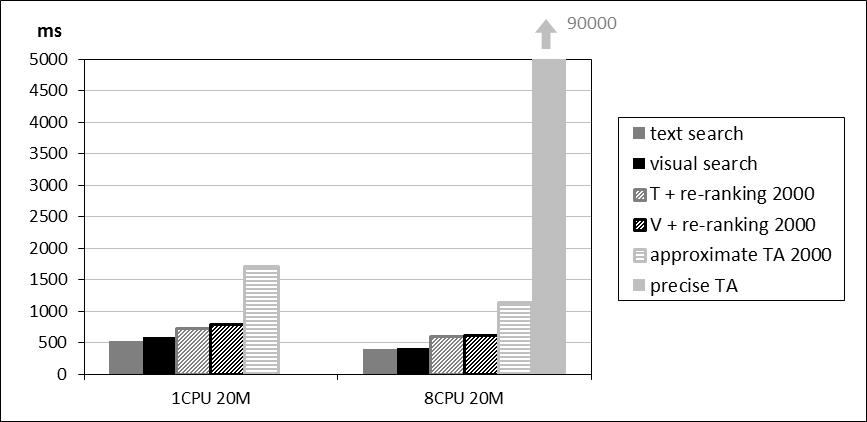


Figure 5: Average time of various fusion method evaluations.

The averaged response times for the monitored fusion techniques can be seen in Figure 5. We can see that the times for the two baseline single-modal searches (visual and text) are increased in the postprocessing phase by approximately 200-300 milliseconds (which represents about 30 % increase) in all cases. This represents the time needed to pass the candidate set to the ranking phase and compute the combined distances. Quite noticeable are the high costs of the approximate TA that are more than two times higher than in case of postprocessing. This is caused by the need to access two index structures, which results in increased communication costs. The indexes also compete for the single machine resources. This is improved as the parallelization is increased using more CPUs but still the method is nearly two times slower than the asymmetric postprocessing fusion. Out of the scope of the graph is the time of the precise TA that took about 1.5 minutes to compute on average, which is caused mainly by the fact that the ordered lists of candidates needed to be examined very deeply before the precise stopping condition was satisfied.

Inherent fusion performance

The inherent fusion technique was proposed to support asymmetric late fusion in a highly flexible and efficient way, allowing the search system to evaluate a significant portion of the database with respect to both modalities. In this section, we analyze its effectiveness and efficiency. For comparison, we also consider the performance of the re-ranking methods and the precise TA, which represent the theoretical lower and upper bound, respectively, on results quality as well as processing costs.

Distance-based result quality

Figure 6 plots the *rED* curves of different re-ranking and inherent fusion methods for both visual-based and text-based searching. We can observe that for all asymmetric processing methods, the error continues to decrease with the growing size of  and . However, it is important to notice that the dependence between the objective quality and the initial result set size is approximately logarithmic. This is clearly visible for visual-based approaches, which we tested with more values of super-candidate set size . For text-based solutions,  larger than 30000 would not bring noticeable improvements, as there are not enough objects relevant from the text perspective that could enter the fusion phase. Even for the 30000 limit, about 40 % of our queries did not have that many text candidates.

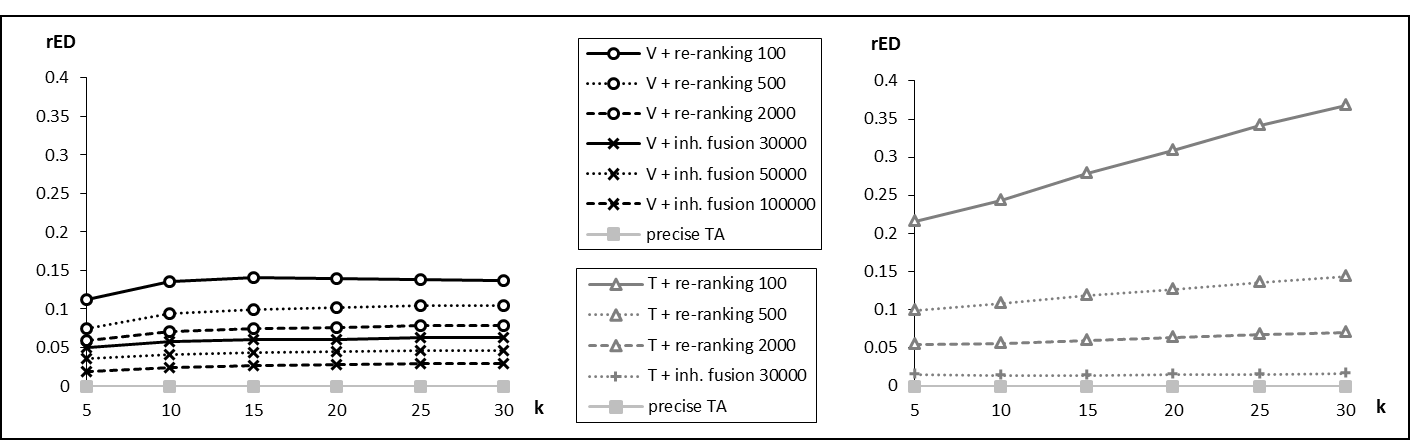


Figure 6: Inherent fusion: Average distances at a given rank for visual primary modality (left) and text primary modality (right).

User-perceived quality

The NDCG quality measure confirms that the inherent fusion technique improves result quality (see Figure 7). Interestingly, the text-based search with inherent fusion outperforms even the precise Threshold Algorithm. This is again caused by the fact that users are more tolerant towards objects that are semantically relevant and less visually similar than to the inverted case. We should also notice that Figure 7 displays the natural NDCG mode. In the strict mode, the relevance achieved by text with inherent fusion is about the same as for TA.

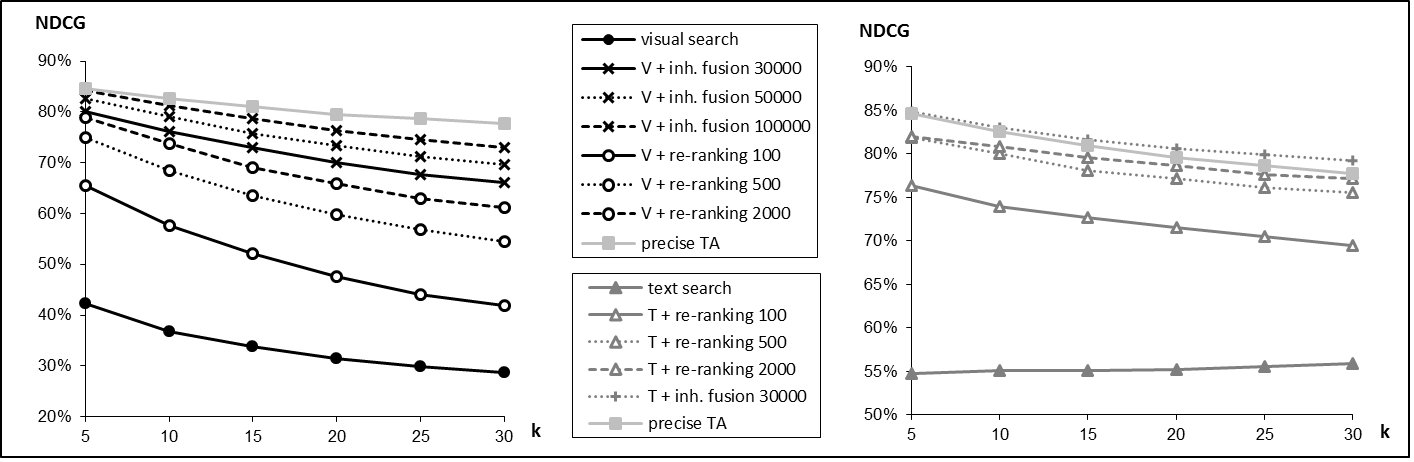


Figure 7: Inherent fusion: Average NDCG at a given rank for visual primary modality (left) and text primary modality (right).

Efficiency

Similarly to the previous section that compared the efficiency of the late fusion techniques, we have also measured the wall-clock time for the inherent fusion method. The average costs of all inherent fusion methods are summarized in Table 1. Comparing the values to those in Figure 5, we can clearly see that inherent fusion on 30000 objects outperforms (in terms of processing time) all the postprocessing methods which process much smaller candidate set. This is caused by the fact that the re-ranking methods need to wait for the index to supply the full result before the ranking is computed. Naturally, the costs of inherent fusion increase with the growing number of objects that are processed. Still, the overall processing time remains acceptable even for the inherent fusion on 100000 objects, especially in the multi-CPU setting.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | T + inh. fusion 30000 | V + inh. fusion 30000 | V + inh. fusion 50000 | V + inh. fusion 100000 |
| 1 CPU | 632 ms | 677 ms | 852 ms | 1224 ms |
| 8 CPU | 461 ms | 489 ms | 603 ms | 826 ms |

Table 1: Inherent fusion: Average time of various fusion method evaluations.

Efficient and effective solutions for multi-modal image retrieval

Having analyzed the general behavior of commonly used late fusion solutions and the newly proposed inherent fusion technique, we can now focus more closely on the methods that have been shown to be efficient enough in large-scale environment. In particular, we are interested in discovering the limitations of their applicability and the potentials for further search quality improvement.

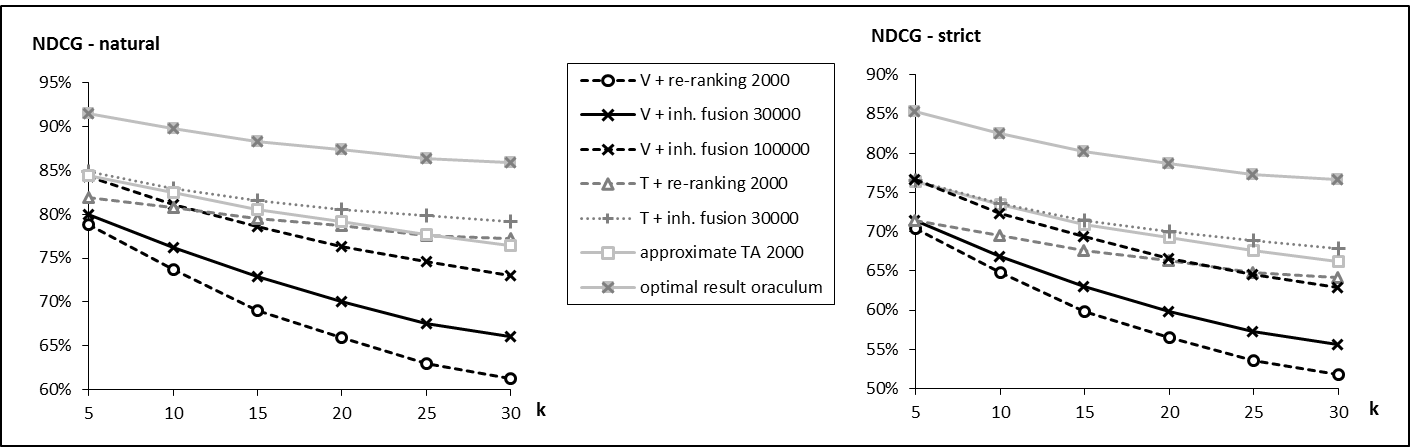


Figure 8: Selected late fusion methods: Average NDCG at a given rank.

Figure 8 summarizes the user-perceived retrieval quality of methods that are applicable for interactive large-scale searching. The inherent fusion techniques together with approximate TA clearly dominate the graph. Since the costs of approximate TA are several times larger than for the inherent fusion, the latter is the more eligible solution. Deciding upon our overall results, the text-based retrieval with inherent fusion is the optimal method for the given dataset.

If we leave the average performance values and focus on the relevance of results for individual query objects, we discover that the dominance of text-based fusion is the most stable from the available options (Figure 9), but not ubiquitous. Actually, the text-based inherent fusion results are best only for approximately 20  % of queries. If we were able to guess which retrieval method is best suited for which query, we could increase the average result relevance by 10  % as depicted in Figure 8 by the “optimal result oraculum” line. Obviously, deciding the suitability of a given retrieval method for a given query is a very challenging task and it remains open for future study. Currently, we have been able to identify several categories of queries for which the text-based asymmetric approaches are less suitable than visual-based ones: complex queries (“two coins”, illustrated in Figure 9), ambiguous queries (“shells”, “stamp”), and too broad queries (“bird”). In our future work, we would like to focus on determining common characteristics of objects in these classes that would allow us to automatically recognize queries that should be processed by visual-based asymmetric fusion.

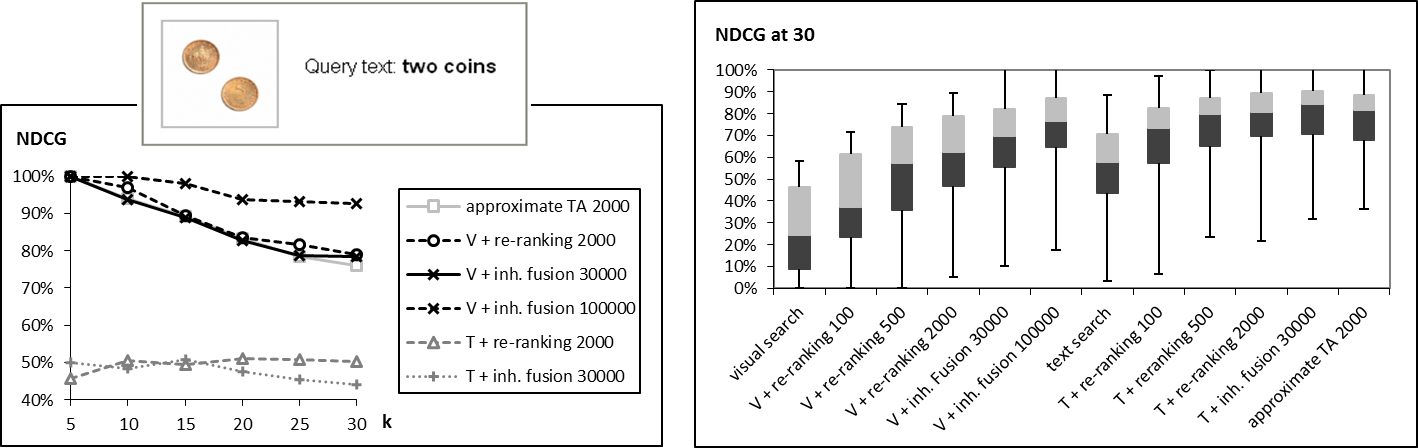


Figure 9: Selected late fusion methods: Stability of NDCG (left), NDCG for a specific query (right).

Conclusions and future work

In this paper, we address several aspects of multi-modal search methods for large-scale image retrieval. Reflecting the rapid development of multi-modal searching and the need for comparative evaluations of different approaches, we provide a theoretical analysis and experimental verification of performance of selected late fusion techniques. In addition to the standard solutions, which are identified and described in the first part of this work, we present a novel technique of inherent fusion, which is shown to achieve a very good performance-costs trade-off over real-world data. Using this technique, we are also able to provide a scalable multi-modal solution, since the scalability of underlying index structure is straightforwardly exploited.

In the second half of the paper, we analyze the results of a comprehensive experimental evaluation of all discussed methods over a large collection of real-world image data. We study the theoretical precision as well as user-satisfaction and compare these with the processing costs of individual methods. The experimental results enable us to select eligible solutions for the task of web-like image searching, and to identify some limitations of these methods that outline directions for future improvements.

Experimental results presented in this work were obtained by evaluating the search methods over a high-quality image collection. Naturally, such testbed is relevant only for a portion of real-world applications. To provide a more comprehensive view on the performance of different retrieval solutions and their applicability to various search scenarios, we are currently preparing another evaluation of the same set of experiments over a dataset with low-quality text data.

References

Atrey, P. K., Hossain, M. A., El-Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubsky, J., and Zezula, P. (2010). Building a web-scale image similarity search system. *Multimedia Tools and Applications*, 47:599–629.

Batko, M., Kohoutkova, P., and Zezula, P. (2008). Combining metric features in large collections. In *24th International Conference on Data Engineering Workshops (ICDE 2008)*, pages 370–377. IEEE Computer Society.

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

Bozzon, A. and Fraternali, P. (2010). Chapter 8: Multimedia and multimodal information retrieval. In *Search Computing*, volume 5950 of *LNCS*, pages 135–155. Springer Berlin / Heidelberg.

Budikova, P., Batko, M., and Zezula, P. (2011). Evaluation platform for content-based image retrieval systems. In *International Conference on Theory and Practice of Digital Libraries*, pages 1–12.

Budikova, P., Batko, M., and Zezula, P. (2012). Multi-modal image search for large-scale applications. In *International Workshop on Multimedia Databases and Data Engineering*, pages 1–7.

Chatzichristofis, S. A., Zagoris, K., Boutalis, Y. S., and Arampatzis, A. (2012). A fuzzy rank-based late fusion method for image retrieval. In *18th International Conference on Advances in Multimedia Modeling (MMM 2012)*, pages 463–472.

Chen, J., Ma, R., and Su, Z. (2010). Weighting visual features with pseudo relevance feedback for CBIR. In *9th ACM International Conference on Image and Video Retrieval (CIVR 2010)*, pages 220–227. ACM.

Cong, G., Jensen, C. S., and Wu, D. (2009). Efficient retrieval of the top-k most relevant spatial web objects. *Proceedings of the VLDB Endowment*, 2(1):337–348.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40:5:1–5:60.

Depeursinge, A. and Müller, H. (2010). Fusion techniques for combining textual and visual information retrieval. In *ImageCLEF*, volume 32 of *The Kluwer International Series on Information Retrieval*, pages 95–114. Springer Berlin Heidelberg.

Fagin, R. (2002). Combining fuzzy information: an overview. *SIGMOD Record*, 31:109–118.

Hörster, E., Slaney, M., Ranzato, M., and Weinberger, K. (2009). Unsupervised image ranking. In *1st ACM workshop on Large-scale multimedia retrieval and mining (LS-MMRM ’09)*, pages 81–88, New York, NY, USA. ACM.

Jain, R. and Sinha, P. (2010). Content without context is meaningless. In *Proceedings of the international conference on Multimedia (MM ’10)*, pages 1259–1268, New York, NY, USA. ACM.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Jegou, H., Schmid, C., Harzallah, H., and Verbeek, J. J. (2010). Accurate image search using the contextual dissimilarity measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):2–11.

Jing, Y. and Baluja, S. (2008). VisualRank: Applying PageRank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1877–1890.

Kherfi, M. L., Ziou, D., and Bernardi, A. (2004). Image retrieval from the World Wide Web: Issues, techniques, and systems. *ACM Computing Surveys*, 36:35–67.

Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2:1–19.

Liu, Y., Mei, T., and Hua, X.-S. (2009). CrowdReranking: exploring multiple search engines for visual search reranking. In *SIGIR*, pages 500–507. ACM.

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

McCandless, M., Hatcher, E., and Gospodnetic, O. (2010). *Lucene in Action*. Manning Publications.

Mironica, I., Ionescu, B., and Vertan, C. (2012). Hierarchical clustering relevance feedback for content-based image retrieval. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012)*, pages 1–6.

MPEG-7 (2002). Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002.

Novak, D., Batko, M., and Zezula, P. (2011). Metric index: An efficient and scalable solution for precise and approximate similarity search. *Information Systems*, 36(4):721–733.

Wang, L., Yang, L., and Tian, X. (2009). Query aware visual similarity propagation for image search reranking. In *17th International Conference on Multimedia (ACM Multimedia 2009)*, pages 725–728.

Zezula, P., Amato, G., Dohnal, V., and Batko, M. (2006). *Similarity Search - The Metric Space Approach*, volume 32. Springer.

Zitouni, H., Sevil, S. G., Ozkan, D., and Duygulu, P. (2008). Re-ranking of web image search results using a graph algorithm. In *19th International Conference on Pattern Recognition (ICPR 2008)*, pages 1–4. IEEE.