Journal of Database Management
Volume 27 • Issue 4 • October-December 2016

Inherent Fusion:
Towards Scalable Multi-Modal Similarity Search

Petra Budikova, Masaryk University, Brno, Czech Republic
Michal Batko, Masaryk University, Brno, Czech Republic
David Novak, Masaryk University, Brno, Czech Republic
Pavel Zezula, Masaryk University, Brno, Czech Republic

ABSTRACT

The rapid growth of unstructured data, commonly denoted as the Big Data challenge, requires new technologies that are capable of dealing with complex data objects such as multimedia. In this work, the authors focus on the content-based retrieval approach, which is able to organize such data by exploiting the similarity of data content. In particular, they focus on solutions that are able to combine multiple similarity measures during the query evaluation. The authors introduce a classification of existing approaches and analyze their performance in terms of effectiveness, efficiency, and scalability. Further, they 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 discussed methods is evaluated by extensive experiments with user participation.*

KEYWORDS

Content-Based Retrieval, Evaluation, Image Retrieval, Late Fusion, Multi-Modal Search, Scalability, Similarity Searching

INTRODUCTION

Due to the rapid growth of both the amount and the diversity of digital data, many organizations are nowadays facing the Big Data problem, i.e. the situation when they have potential access to a wealth of information, but they do not know how to get value out of it (Zikopoulos & Eaton, 2011). This inspired research and new solutions on different levels of data processing, including data modeling, storing, and analysis (Batini et al., 2015; Maté et al., 2015; Meged & Gelbard, 2012). In this paper, we focus on the storage and retrieval of unstructured data that cannot be straightforwardly organized in relational databases, since they are not searched by exact match but rather by similarity. This is e.g. the case of images, where pixel-to-pixel matching does not make sense but searching by visual similarity is desired in many situations, e.g. medical image analysis, entertainment, security, and surveillance. For such applications, the content-based retrieval techniques were developed (Alemu et al., 2009; Datta et al., 2008).

The fundamental idea of content-based data management is to organize complex data objects such as multimedia using their *content* instead of *descriptive metadata* that are used in traditional data management systems. As illustrated in Figure 1, a content-based image search system can retrieve images that are visually similar to a given example. A principal advantage of the content-based paradigm is the fact that the multimedia object content is always available, whereas the metadata are often sparse, erroneous, or not available at all. Depending on the type of the data to be processed,

DOI: 10.4018/JDM.2016100101

different salient features can be extracted from the complex objects and used for indexing and retrieval of the original data. In case of images, we can use e.g. global features such as MPEG7 color, shape, or texture, local image features describing individual points of interest, face descriptors, etc. The relevance of individual data items with respect to a given query is then determined by the similarity of the extracted features, which is computed by a suitable distance function (Zezula et al., 2006). In this paper, we shall call each salient feature and the associated distance function a *modality* of the content-based similarity search.

In the first content-based multimedia retrieval systems, a single modality was utilized to organize and search the data. However, this proved to be insufficient for several reasons: 1) each modality only reflects a specific perspective of the complex object, which may not agree with the actual users' subjective view (this is often denoted as 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 & Sinha, 2010).

Following these observations, a number of multi-modal retrieval systems have been proposed in the past decade. In this paper, we are mainly interested in image retrieval, in particular a general-purpose image retrieval that could be used e.g. in a web search engine. This task appears in many real-world applications and therefore has attracted many researchers from different communities. As a result, diverse multi-modal image search techniques can be found in the literature. 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.

Figure 1. Content-based image retrieval: similar images to the query (left) were selected from a 20M image collection

Contributions of the Paper

In this paper, we focus on several open issues related to providing efficient and effective solutions for large-scale, general-purpose similarity retrieval. We elaborate on our earlier work (Budikova et al., 2012), where we proposed a simple classification of multi-modal retrieval techniques and experimentally evaluated their basic properties. In the current work, both the classification and the experimental evaluations are more thorough, and we introduce a new technique of inherent fusion. The main contributions of this paper are the following:

- Survey and systematic categorization of existing approaches: In the first part of the paper, we formalize the multi-modal retrieval problem and present a categorization of possible approaches. We also discuss the theoretical applicability of selected state-of-the-art solutions.
- Proposal of a novel technique: Next, 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 evaluation: The last part of the paper is devoted to experimental evaluation. Using a uniform 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.

In the analysis of existing approaches and the new technique proposal, we deal with general similarity retrieval that can be applied to various domains. In the experimental section, we focus on a specific task of image retrieval, which is one of the most important applications of content-based searching. Also, sufficient amounts of image data are available so that we can test the scalability of individual methods, and the relevance of visual information is easy to evaluate in user satisfaction experiments.

MULTI-MODAL CONTENT-BASED RETRIEVAL

Before we start analyzing the recent advances in multi-modal searching, let us briefly review the basic techniques of similarity searching and explain more thoroughly the architecture of single-modality and multi-modal retrieval systems.

Similarity Searching

Similarity-based data management is a generic approach that allows to organize and search any data for which a measure of similarity between individual objects can be defined (Zezula et al., 2006). The similarity of objects is typically expressed by the inverse concept of a *distance* (dissimilarity) measured by a suitable distance function. The *distance function* can be applied to any pair of objects from a given domain and produces a positive number or zero; the zero value is returned for identical objects, higher values correspond to a growing dissimilarity between objects. Noticeably, this definition can also accommodate the exact-match paradigm (used in traditional databases) by assigning a fixed non-zero distance to all non-matching object pairs.

Let $X \in D_X$ be a collection of objects to be organized and D_X be the domain of objects from X. The similarity-based data retrieval follows the "query-by-example"' principle, where the query is defined by a reference object $q \in D_X$ and a similarity condition that needs to be satisfied by qualifying objects from X. In this study, we limit our attention to the most typical query type – *the k nearest neighbor (kNN) query*, which retrieves the *k* objects that are most similar to the reference point q. Nearest neighbor queries can be used e.g. to recognize a song from a fragment recording, track objects in videos, or automatically cluster and annotate images. Developing efficient and effective algorithms for kNN queries is thus a very important issue.

Single Modality Retrieval

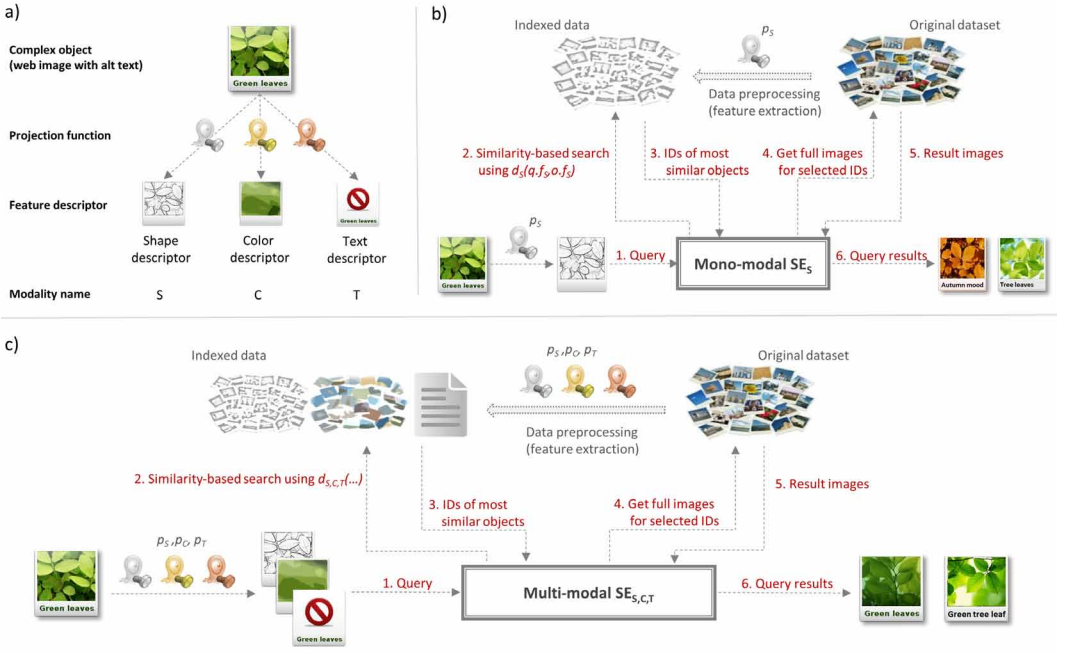
As we already mentioned, complex objects such as multimedia are not indexed and searched in their original form. Instead, salient features are extracted from all objects and used to index and retrieve them. Each type of a feature and its associated distance function are denoted as a *modality* of the content-based retrieval. Informally, a modality can be understood as a point of view, which transforms complex objects from the domain D_X to representations more suitable for searching (see Figure 2a). Formally, we define modality M as an ordered pair $\left(p_M, d_M\right)$, where $p_M : D_X \rightarrow D_M$ is a projection function and $d_M : D_M \times D_M \rightarrow \mathbb{R}_0^+$ is the distance function. The projection function transforms an object $o \in D_X$ into a feature descriptor $o.f_M \in D_M$, while the function d_M evaluates the distance between two descriptors, i.e. the dissimilarity of two objects as seen in the view of the modality M.

A mono-modal search engine SE_S is illustrated in Figure 2b. It employs a shape modality S to organize the data. All objects from the dataset as well as queries are projected into the domain of this modality and all comparisons take into account only shapes. Accordingly, the retrieved images are similar to the query in shape, but may be quite different in other aspects.

Multi-Modal Retrieval

The multi-modal retrieval paradigm assumes the existence of multiple modalities $M_1, \ldots, M_n$, which can be utilized to manage data from *X*. At some point of the data processing, these modalities are combined – *fused* – to provide more complex representations of objects from *X* and to evaluate their similarity on a higher semantic level. The fusion can be realized by a multi-modal projection function, multi-modal distance function, or both. A multi-modal projection function $p_{\widehat{M_i,\ldots,M_j}}$ typically takes into account the outcomes of original mono-modal functions $p_{M_i}, \ldots, p_{M_j}$ and combines these into a

Figure 2. Image search example: a) Modalities, b) single-modality retrieval, c) multimodal retrieval



more complex object descriptor, whereas a multi-modal distance function $d_{\widetilde{M_i,\ldots,M_j}}$ mostly aggregates mono-modal distances produced by $d_{M_i},\ldots,d_{M_j}$. A kNN query over $SE_{M_1,\ldots,M_n}$ is then defined by a multi-modal query object *q* and a multi-modal query distance function d_Q.

A multimodal search system is depicted in Figure 2c. It takes into account three modalities of each image – color, shape, and text description. As a result, a different set of images is returned as the most similar objects. In this schema, there is a separate index for each of the three modalities, which corresponds to a specific type of late fusion. Basically, there are two ways to design a multi-modal system:

- *Early fusion* approaches combine modalities $M_1,\ldots,M_n$ 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 (Atrey et al., 2012; Depeursinge & Müller, 2010; Escalante et al., 2012; Tran et al., 2013). The actual indexing and retrieval then work the same way as in a mono-modal system.
- *Late fusion* techniques, on the other hand, fuse 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 d_Q (Atrey et al., 2012; Chatzichristofis et al., 2012; Depeursinge &Müller, 2010; Fagin, 2002; Jing & Baluja, 2008; Wang et al., 2009).

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 – since the understanding of similarity is known to be subjective and context-dependent, the system should allow users to dynamically adjust the similarity measure d_Q to their individual needs. Early fusion techniques typically allow to search data only by the distance function that was chosen during the data indexing phase, and therefore are not suitable for such tasks. Late fusion in principle does not pose any restrictions on d_Q and is thus better suited for adaptive searching. Therefore, we limit our discussions to the late fusion approach in the following analysis.

LATE FUSION TECHNIQUES: CLASSIFICATION AND STATE-OF-THE-ART

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 & Fraternali, 2010; Depeursinge & Müller, 2010) to get a more complex view of the problem.

The objective of this paper is to analyze late fusion methods and assess their suitability for large-scale image retrieval. For this purpose, we first introduce two criteria that are important for scalable searching and can be used to classify existing solutions. Then we present the most important works in each class and discuss their strengths and weaknesses.

Classification of Late Fusion Methods

Late fusion methods can be 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 processed in the fusion phase. In this section, we explore each of these aspects in more detail, focusing on the theoretical fusion principles. Specific techniques implementing individual fusion paradigms will then be discussed in the next section.

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 (e.g. color-based and shape-based distance are evaluated separately and combined later), or candidate sets selected by mono-modal index structures (e.g. we first select a candidate set using shape modality and then apply further criteria). 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 $M_1,\ldots,M_n$ are exploited for parallel retrieval of candidate sets $C_1,\ldots,C_n$, which are then merged into a single candidate set *C* and re-evaluated with respect to the query distance d_Q (see Figure 3-left). All modalities are exploited at the same time and with the same level of importance – the d_Q 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 M^P) and used to retrieve the candidate set *C*, which is further processed using d_Q (see Figure 3-right). The remaining (secondary) modalities, denoted as M^S, are less influential than the primary ones – an object that does not rank high with the primary modalities will not appear in *C*, even if it would have been considered perfect by some of the secondary modalities.

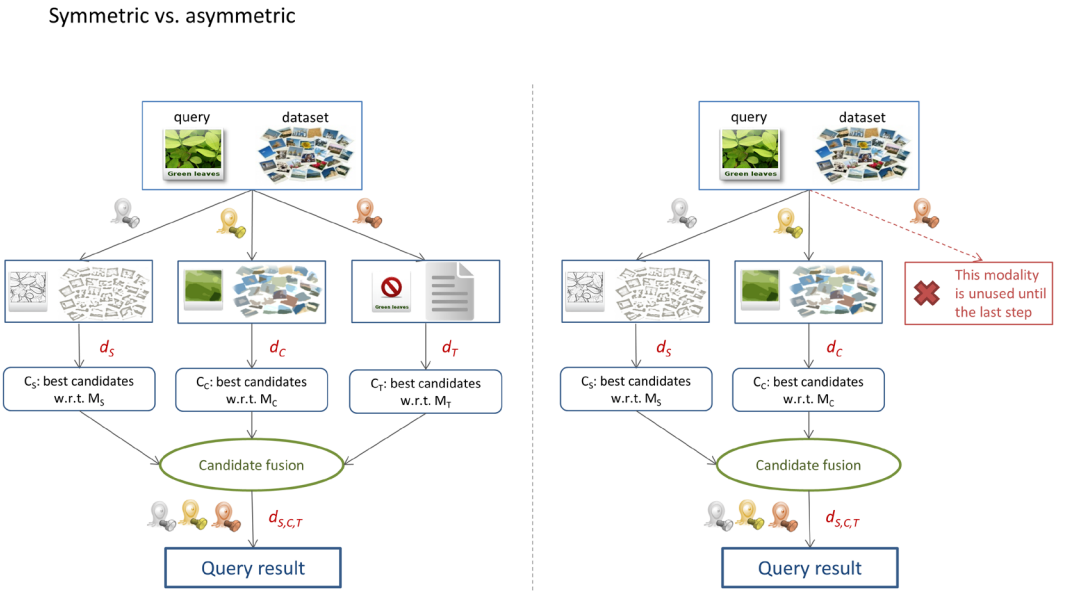
The reason for the utilization of asymmetric solutions may be threefold: 1) the primary modalities may be more vital for a given use-case scenario; 2) the asymmetric solution may be chosen because of efficiency issues; or 3) some of the modalities may not be available at the beginning of the query evaluation. We shall explore the latter two reasons in more detail in the following sections.

Fusion Scenarios

The second classification aspect we would like to discuss concerns the selection of objects for which the multi-modal distance d_Q is evaluated. In the late fusion paradigm, objects from dataset *X* are organized in one or more index structures, where each index exploits just one modality. These mono-modal indexes provide candidate objects relevant for the particular modality, and the multi-modal distance is then computed for the candidates. The quality of the retrieval results as well as the query evaluation costs are closely related to the strategy of selecting the candidate objects for the actual modality fusion.

We find it helpful to separate late fusion strategies into two classes according to the phase of query evaluation in which the fusion takes place. In general, the similarity query evaluation can consist of multiple phases, including query preprocessing, basic search, result postprocessing, presentation, and possibly a user-feedback loop (see Figure 4). Among these, the late fusion can take place either in the basic search, or during result postprocessing:

Figure 3. Symmetric fusion (left): all modalities are utilized during the whole query processing; asymmetric fusion (right): color and shape features are primary and are utilized during the whole processing, whereas the secondary text modality is considered only for the candidate set selected by the primary modalities.

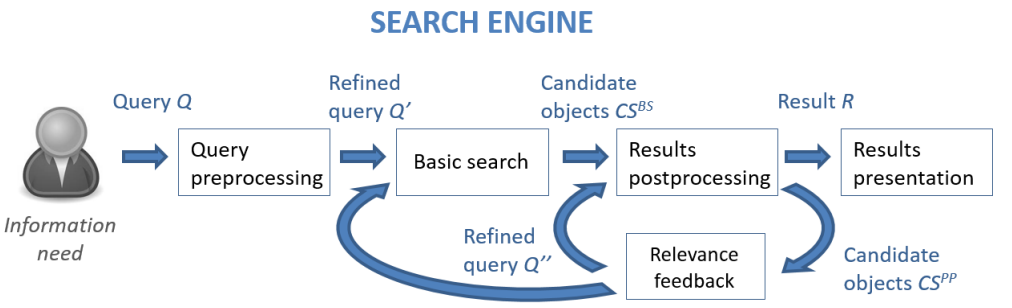


- *Basic search phase* is the core part of any query evaluation. It accesses index structures and identifies candidate objects. Processes evaluated in the basic search phase have access to all objects from the dataset *X*. We classify a fusion technique as a basic-search fusion, if the technique allows to access all objects that are needed to provide the best possible results with respect to d_Q.
- *Postprocessing phase* follows after the basic search and evaluates additional computations over a set of candidate objects C^{BS}, 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 C^{BS} – 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, C^{BS} is typically much smaller than the original dataset *X*, therefore the fusion costs are relatively low and independent of the dataset size. Moreover, the postprocessing fusion may exploit the properties of candidate objects in C^{BS} 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 Late Fusion Techniques

In the previous section, we have introduced two aspects for the classification of late fusion methods. In each aspect, two types of solutions were defined, which results in four categories of late fusion techniques. Let us now populate these categories by state-of-the-art methods 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.

Figure 4. Phases of query evaluation



Symmetric Fusion in Basic Search Phase

Symmetric late fusion in the basic search phase is best represented by the Threshold Algorithm, which was introduced by Ronald Fagin (Fagin, 2002). Candidate objects 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 $M_1,\ldots,M_n$ be the input modalities and $L_1,\ldots,L_n$ be ordered lists of objects from X defined by individual modalities. In each iteration, the Threshold Algorithm takes the top unvisited object from each sorted list, adds it to candidate set C, and evaluates its overall distance d_Q from the query object q. Then, a threshold condition is verified which decides whether the top k objects in C represent the best possible result, or another iteration needs to be executed.

The Threshold Algorithm represents a theoretically sound, clear solution that provides a precise answer and 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 would 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 X is evaluated with respect to all modalities, but only a subset of the 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 two specific data modalities – text description and location. To the best of our knowledge, similar solutions only exist for the geo-textual retrieval. 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 L_i provided by modality M_i. 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 *X*.

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, Batko et al. 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 & 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 similarity-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 C^{BS}, 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 C^{BS} is typically updated in the postprocessing and the top-ranked 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 makes it possible to use relevance feedback (RF) and pseudo-RF processing.

Ranking by orthogonal modality is well known from commercial image search systems Google (Jing & 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 (Li et al., 2012), which proposes several techniques for textual ranking of C^{BS} and discusses pseudo-RF ranking techniques that exploit information learned from objects in C^{BS}. Pseudo-RF postprocessing is also utilized in (Chen et al., 2010) to adjust the weights of multiple visual features that are fused. Solutions presented in (Hörster et al., 2009; Mironica et al., 2012; Zitouni et al., 2008) 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) proposes 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 indexing 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 suitable primary modalities are available in sufficient quality.

THE INHERENT FUSION TECHNIQUE

The key limitation of the asymmetric postprocessing fusion is the fact that the candidate set C^{BS} passed to the multi-modal ranking phase has a given size. The larger is this candidate set, the greater is the probability that the most query relevant objects will be found, but then 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 candidate set C^{BS} is small, the overall result will be strongly affected by the primary modality, since the objects that would be ranked high by the secondary modality (but not by the primary one) are unlikely to appear among the candidates.

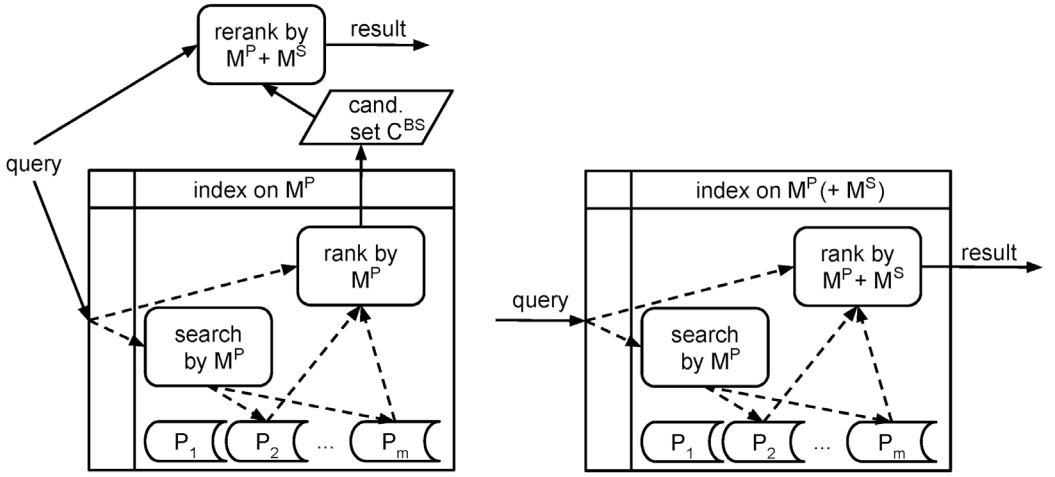
We believe that it is possible to significantly improve the performance of these asymmetric solutions, if we more thoroughly exploit all information available during the query evaluation process. Let us first have a closer look at the processing in the basic search phase. In general, indexing techniques typically partition the dataset into a number of, not necessarily disjoint, "data chunks" (intervals, areas, clusters, posting lists, etc.); let us denote these partitions $P_1,\ldots,P_m$, where $P_i \subseteq X$. During evaluation of query kNN(q), the search algorithm identifies some of these partitions that could potentially contain data relevant to query q with respect to the primary modalities M^P; let us denote objects from these selected partitions as a "primary candidate set" C^P. Objects from this set are actually accessed during the search process and the best objects according to M^P then form the candidate set C^{BS} for postprocessing. The size of primary candidate set C^P is usually significantly larger (orders of magnitude) than the typical size of C^{BS}.

Following this observation, we propose a novel technique of *inherent fusion* that utilizes all objects visited by the index also for ranking by secondary modalities. Similar to the asymmetric fusion techniques, we propose to index the data using selected primary modalities M^P. However, the full data objects are stored, so that all primary and secondary modalities (M^P and M^S) are held by the index. At query time, the retrieval algorithm accesses objects from the primary candidate set C^P but it does not rank them with respect to the primary modality distance function d_{M^P}, but each object from C^P is ranked by the multi-modal similarity function d_Q that 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 5. In case of the postprocessing (left schema), the index on M^P identifies relevant partitions P_i and ranks the data from these partitions by d_{M^P} to create the candidate set for further processing by the multi-modal query distance d_Q. As described above, the inherent fusion (Figure 5, right) ranks the objects directly by d_Q as they are accessed. In comparison with the postprocessing fusion, the volume of data searched with all modalities $|C^P|$ is considerably larger, increasing the probability of discovering more relevant objects. Importantly, this processing is far less costly than the standard asymmetric postprocessing fusion on a candidate set of size $|C^P|$ because that would typically require processing of even larger C^P within the M^P index.

The inherent fusion can be implemented relatively easily within most of the standard indexing techniques if several adaptations are made. First, the index needs to store the data for the secondary

Figure 5. Schema of asymmetric postprocessing fusion (left) and inherent fusion (right)



modalities M^S so that the ranking can be applied, but storing additional 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 M^P and M^S are combined at query time (for instance, arbitrary weighting) and there are no limitations on the indexability of the additional modalities (M^S);
- *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 C^S 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 somewhere in any case for the postprocessing phase so the only difference is in the implementation of the storage facilities.

EXPERIMENTAL EVALUATION

The eligibility of any search method is strongly influenced by two natural quality measures – its computational efficiency and relevance of the 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, evaluation of the result quality 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 to 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 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 have 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 & Ribeiro-Neto, 2011), whereas the visual similarity is evaluated by a static combination of selected MPEG-7 descriptors (Batko et al., 2008). These two modalities are fused via distance aggregation function that will be discussed in more detail later. As for the search methods, 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 C^{BS};
- 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 C^{BS} and C^{P} 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 6 shows a few queries from our selection.

Ground Truth

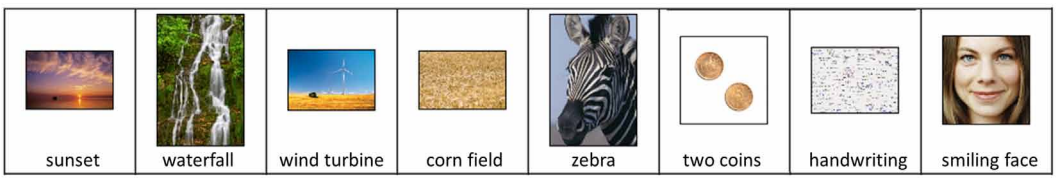
To be able to evaluate user-perceived relevance of retrieved objects, it was further necessary to provide ground truth data for our queries. Since we could not hope to obtain the ground truth assessments for all 20M Profiset images, we collected a *partial ground truth* in the following way: for each query, top-30 queries were run using each of the methods and each 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. The partial ground truth that was obtained in this way 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.

EXPERIMENTAL RESULTS AND DISCUSSION

In the above described experiments and during the ground truth collection process, we acquired large amounts of real-use observations 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:

Figure 6. Query objects



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: 1) the objective result quality, as measured by the distance function d_Q; 2) the subjective result quality as perceived by users; and 3) 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 arbitrarily, 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 & 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 $d_Q\left(q,o\right)=c_{Vnorm}\times d_V\left(p_V\left(q\right),p_V\left(o\right)\right)+c_{Tnorm}\times d_T\left(p_T\left(q\right),p_T\left(o\right)\right)$ is considered in all following comparisons.

Comparative Analysis of State-of-the-Art Late Fusion Techniques

In the first set of evaluations, we compare the performance of standard late fusion methods (not including the inherent fusion). The Threshold Algorithm (TA) represents the only solution that guarantees a precise result, 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 using different primary modalities 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 R^A to a precise retrieval result R^P. 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 the *k*-th position (d^k) in approximate and precise results: $rED\left(k\right)_{R^A}=d^k_{R^A}\,/\,d^k_{R^P}-1$.

The results provided by the *rED*(*k*) measure are depicted in Figure 7. 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 C^{BS} which enters the postprocessing (fusion) phase. We can notice that text-based initial search followed by a re-ranking of small C^{BS} 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 C^{BS}, the retrieval accuracy gradually improves for all approximate techniques. For the comparable size of the candidate set $|C^{BS}| = 2000$ 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.

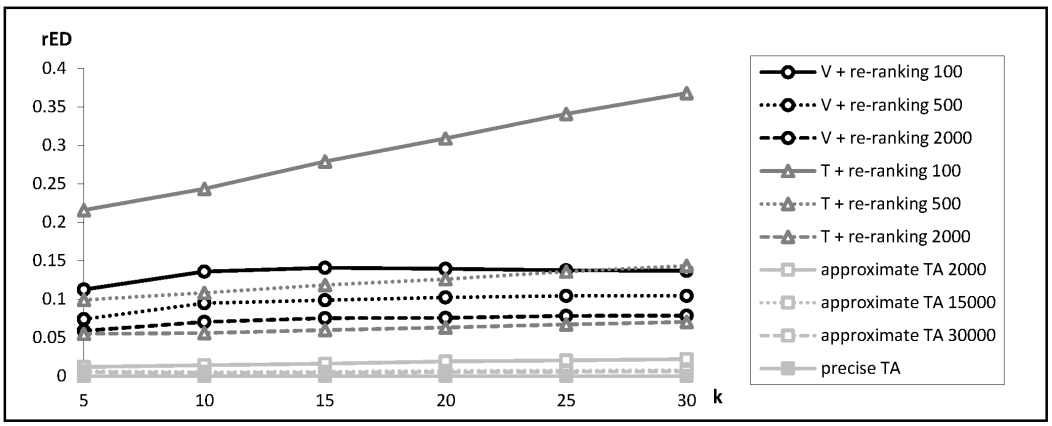
User-Perceived Quality

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 & 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 *partially relevant* during user relevance assessments are considered to have non-zero relevance, whereas in the *strict* mode, only objects that were marked as *highly relevant* are deemed relevant. The strict measure thus represents a more demanding user.

In Figure 8 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. The users were better satisfied by the results provided by the text-based 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 interesting that the relevance of results continues to increase with the growing size of C^{BS}. 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 $|C^{BS}|$ was exceeded. However, these solutions only applied the secondary modality to determine the ranking in

Figure 7. Standard late fusion: Average relative error on distance at a given rank



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 retrieval 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 second set of experiments we have used all 8 CPUs and thus allowed the indexes to utilize their internal parallelization.

The averaged response times for the monitored fusion techniques can be seen in Figure 9. We can observe 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 Evaluation

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 primary and secondary modalities. In this section, we analyze the effectiveness and efficiency of inherent fusion. 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.

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

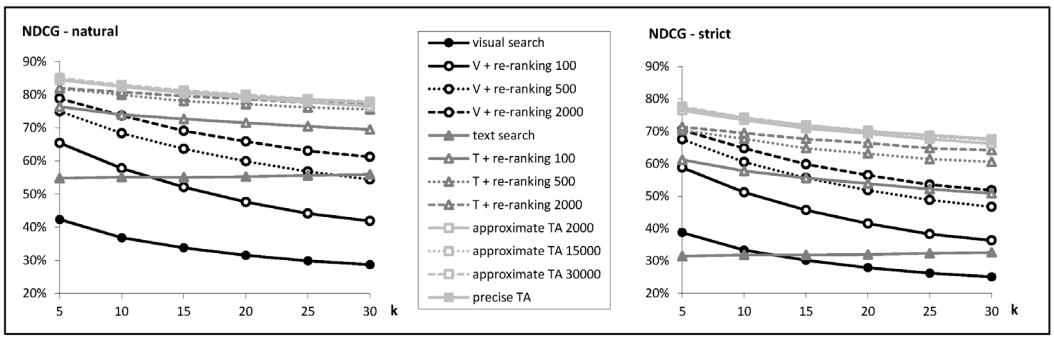
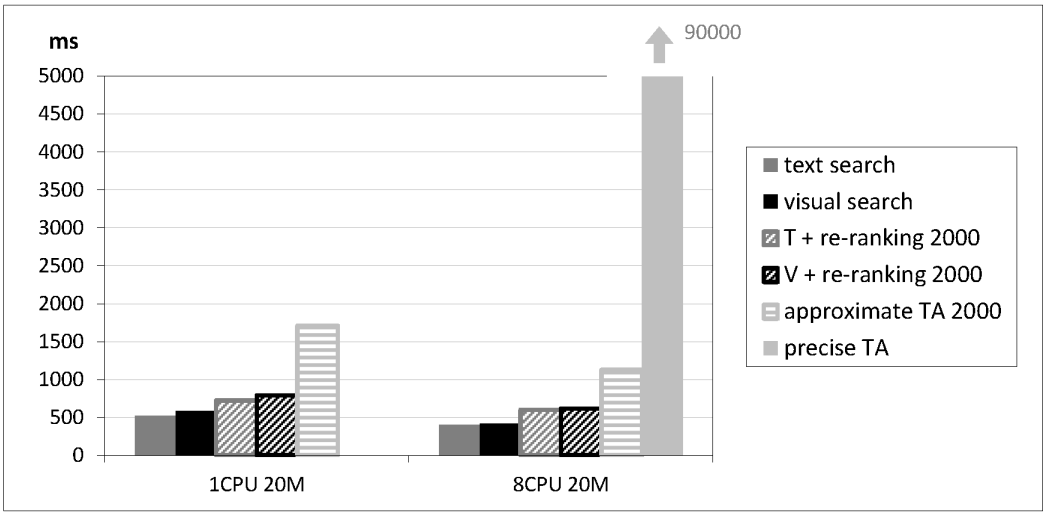


Figure 9. Average time of various fusion method evaluations



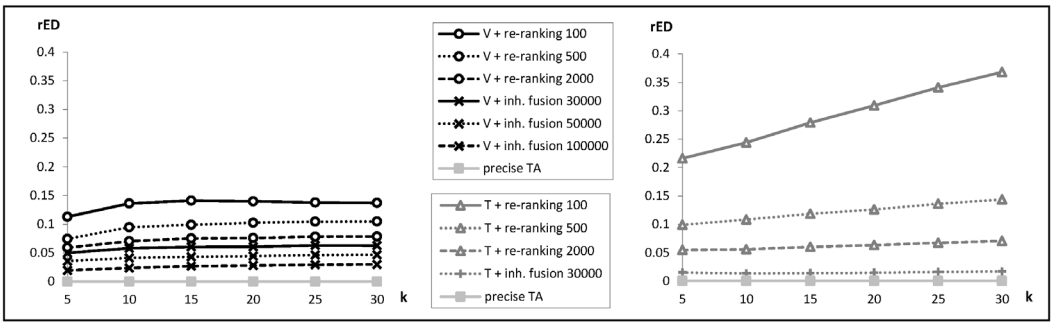
Distance-Based Result Quality

Figure 10 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 C^{BS} and C^{P}. 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 primary candidate set size $|C^{P}|$. For text-based solutions, $|C^{P}|$ larger than 30,000 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 30,000 limit, about 40% of our queries did not have that many text candidates.

User-Perceived Quality

The NDCG quality measure confirms that the inherent fusion technique improves result quality (see Figure 11). 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

Figure 10. Inherent fusion: Average relative error on distance at a given rank for visual primary modality (left) and text primary modality (right).



that are semantically relevant and less visually similar than to the inverted case. We should also notice that Figure 11 displays the natural NDCG mode. In the strict mode, the relevance achieved by text with inherent fusion is about the same as for TA.

Efficiency

Similarly to the previous section that compared the efficiency of the standard 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 9, we can clearly see that inherent fusion on 30,000 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 100,000 objects, especially in the multi-CPU setting.

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 potential for further search quality improvement.

Figure 12 summarizes the user-perceived retrieval quality of methods that are efficient enough to be 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 focus on the relevance of results for individual query objects, we discover that the dominance of text-based fusion is the most stable one among the available options (Figure 13), but not ubiquitous. Actually, the text-based inherent fusion results are the best only for approximately 20 % of queries.

Figure 11. Inherent fusion: Average NDCG (natural) at a given rank for visual primary modality (left) and text primary modality (right).

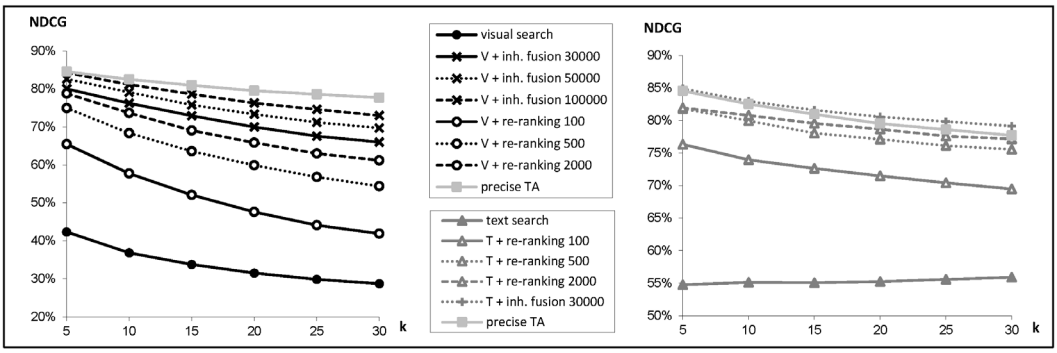


Table 1. Inherent fusion: Average time of various fusion method evaluations

	T + inh. fusion 30,000	V + inh. fusion 30,000	V + inh. fusion 50,000	V + inh. fusion 100,000
1 CPU	632 ms	677 ms	852 ms	1224 ms
8 CPU	461 ms	489 ms	603 ms	826 ms

Figure 12. Selected late fusion methods: Average NDCG at a given rank

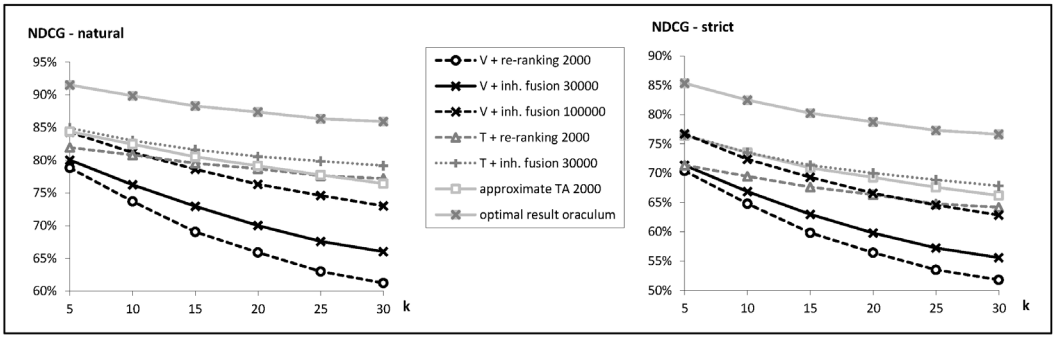
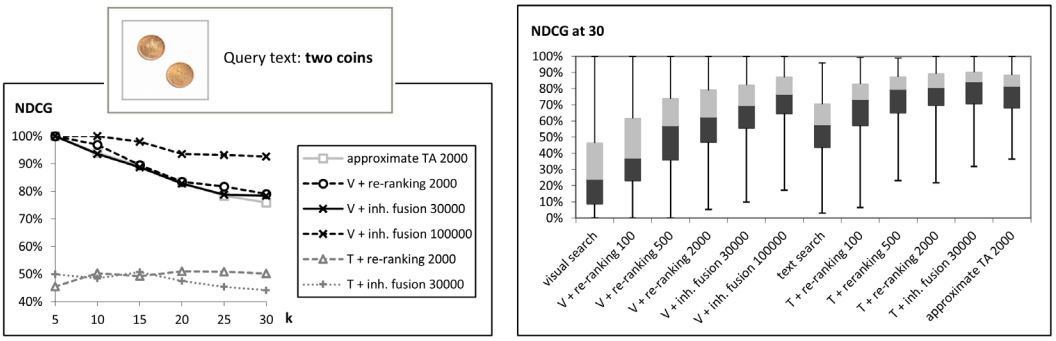


Figure 13. Selected late fusion methods: NDCG for a specific query (left), stability of NDCG (right)



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 12 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 13), ambiguous queries ("shells", "stamp"), and queries with a large variability of visual representations ("bird"). In our future work, we would like to focus on determining common characteristics of objects in these classes, which would allow us to automatically recognize queries that should be processed by visual-based asymmetric fusion.

LESSONS LEARNED AND PRACTICAL IMPLICATIONS

In this paper, we have addressed several aspects of multi-modal search methods for large-scale similarity retrieval. On the theoretical level, we have analyzed existing late fusion techniques and their applicability to large-scale searching. Further, we have proposed a new technique of inherent fusion. A practical contribution of our work is a comparison of diverse late fusion methods on an equal ground and with user feedback. The most important findings of our research can be summarized as follows:

- We have confirmed that the utilization of multiple modalities improves the quality of content-based retrieval, which agrees with the findings of many previous works. However, our experimental comparison of diverse fusion techniques brings new information about the performance of

individual methods on real-world data, which can be very useful for the design of future search systems. In particular, we offer objective measurements of both evaluation costs and results quality of the methods we analyzed.

- For the specific case of image retrieval and the fusion of text and visual modalities, we have observed two generic rules. First, users prefer semantic relevance over visual similarity, therefore text search should be used as the primary modality whenever good quality text metadata is available. Second, the synergy between the two modalities should be exploited as often as possible, therefore a multi-modal distance function should be used in the re-ranking phase instead of a simple visual similarity.
- The proposed inherent fusion technique allows to improve the quality of approximate multi-modal fusion by fully exploiting the data accessed during the query evaluation process. As confirmed by both our theoretical analysis and the experimental verification, the inherent fusion improves the search results without increasing the processing costs. At the same time, the technique is efficient, scalable, and can be easily implemented into existing search systems.

The above-described findings have direct practical implications to the data management and information retrieval fields. With the Big Data phenomenon, multimedia retrieval systems are going to need to adopt the similarity-based searching and combine the content-based modalities with standard descriptive attributes. Already, specific techniques have been proposed that allow to combine similarity-based retrieval with traditional RDBS using extensions of the SQL language (Barioni et al., 2009; Guliato et al., 2009). The presented study can be used as a guide for understanding the modality fusion options and selecting the most suitable fusion strategy for a given application. The new inherent fusion technique is freely available to the data management community within the open-source MESSIF library (Batko et al., 2007).

ACKNOWLEDGMENT

This research was supported by the Czech Science Foundation (GAČR) project No. P103/12/G084.

REFERENCES

Alemu, Y., Koh, J. B., Ikram, M., & Kim, D. K. (2009). Image retrieval in multimedia databases: A survey. *Proceedings of theFifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP'09)* (pp. 681-689). IEEE. doi:10.1109/IIH-MSP.2009.159

Atrey, P. K., Hossain, M. A., El-Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, *16*(6), 345–379. doi:10.1007/s00530-010-0182-0

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search* (2nd ed.). Harlow, England: Pearson Education Ltd.

Barioni, M. C. N., Razente, H. L., Traina, A. J. M., & Traina, C. Jr. (2009). Seamlessly integrating similarity queries in SQL. *Software, Practice & Experience*, *39*(4), 355–384. doi:10.1002/spe.898

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From Data Quality to Big Data Quality. *Journal of Database Management*, *26*(1), 60–82. doi:10.4018/JDM.2015010103

Batko, M., Kohoutkova, P., & Zezula, P. (2008). Combining metric features in large collections. *Proceedings of the24th International Conference on Data Engineering Workshops (ICDE '08)* (pp. 370–377). IEEE Computer Society. doi:10.1109/ICDEW.2008.4498347

Batko, M., Novak, D., & Zezula, P. (2007). MESSIF: Metric similarity search implementation framework. *Proceedings of theFirst International DELOS Conference*, *LNCS* (*Vol. 4877*, pp. 1–10). Springer.

Bozzon, A., & Fraternali, P. (2010). Multimedia and multimodal information retrieval. In Search Computing, LNCS (Vol. 5950, Ch. 8, pp. 135–155). Berlin: Springer.

Budikova, P., Batko, M., & Zezula, P. (2011). Evaluation platform for content-based image retrieval systems. *Proceedings of theInternational Conference on Theory and Practice of Digital Libraries* (pp. 1–12). doi:10.1007/978-3-642-24469-8_15

Budikova, P., Batko, M., & Zezula, P. (2012). Multi-modal image search for large-scale applications. *Proceedings of theInternational Workshop on Multimedia Databases and Data Engineering* (pp. 1–7).

Chatzichristofis, S. A., Zagoris, K., Boutalis, Y. S., & Arampatzis, A. (2012). A fuzzy rank-based late fusion method for image retrieval. *Proceedings of the18th International Conference on Advances in Multimedia Modeling (MMM '12)* (pp. 463–472). doi:10.1007/978-3-642-27355-1_43

Chen, J., Ma, R., & Su, Z. (2010). Weighting visual features with pseudo relevance feedback for CBIR. *Proceedings of the9th ACM International Conference on Image and Video Retrieval (CIVR '10)* (pp. 220–227). ACM. doi:10.1145/1816041.1816075

Cong, G., Jensen, C. S., & Wu, D. (2009). Efficient retrieval of the top-k most relevant spatial web objects. *Proceedings of the VLDB Endowment*, *2*(1), 337–348. doi:10.14778/1687627.1687666

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

Depeursinge, A., & Müller, H. (2010). Fusion techniques for combining textual and visual information retrieval. In ImageCLEF, KISIR (Vol. 32, pp. 95–114). Berlin: Springer. doi:10.1007/978-3-642-15181-1_6

Escalante, H. J., Gómez, M. M., & Sucar, L. E. (2012). Multimodal indexing based on semantic cohesion for image retrieval. *Information Retrieval*, *15*(1), 1–32. doi:10.1007/s10791-011-9170-z

Fagin, R. (2002). Combining fuzzy information: An overview. *SIGMOD Record*, *31*(2), 109–118. doi:10.1145/565117.565143

Guliato, D., de Melo, E. V., Rangayyan, R. M., & Soares, R. C. (2009). PostgreSQL-IE: An image-handling extension for PostgreSQL. *Journal of Digital Imaging*, *22*(2), 149–165. doi:10.1007/s10278-007-9097-5 PMID:18214614

Hörster, E., Slaney, M., Ranzato, M., & Weinberger, K. (2009). Unsupervised image ranking. *Proceedings of the1st ACM workshop on Large-scale multimedia retrieval and mining (LS-MMRM '09)* (pp. 81–88). New York, NY, USA. ACM.

Jain, R., & Sinha, P. (2010). Content without context is meaningless.*Proceedings of the international conference on Multimedia (MM '10)* (pp. 1259–1268).New York, NY, USA. ACM. doi:10.1145/1873951.1874199

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446. doi:10.1145/582415.582418

Jegou, H., Schmid, C., Harzallah, H., & Verbeek, J. J. (2010). Accurate image search using the contextual dissimilarity measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 2–11. doi:10.1109/TPAMI.2008.285 PMID:19926895

Jing, Y., & Baluja, S. (2008). VisualRank: Applying PageRank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(11), 1877–1890. doi:10.1109/TPAMI.2008.121 PMID:18787237

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *2*(1), 1–19. doi:10.1145/1126004.1126005

Li, J., Ma, Q., Asano, Y., & Yoshikawa, M. (2012, April). Re-ranking by multi-modal relevance feedback for content-based social image retrieval. *Proceedings of theAsia-Pacific Web Conference* (pp. 399-410). Berlin: Springer. doi:10.1007/978-3-642-29253-8_34

Liu, Y., Mei, T., & Hua, X.-S. (2009). CrowdReranking: exploring multiple search engines for visual search reranking. Proceedings ofSIGIR (pp. 500–507). ACM. doi:10.1145/1571941.1572027

Maté, A., Llorens, H., de Gregorio, E., Tardío, R., Gil, D., Muñoz-Terol, R., & Trujillo, J. (2015). A Novel Multidimensional Approach to Integrate Big Data in Business Intelligence. *Journal of Database Management*, *26*(2), 14–31. doi:10.4018/JDM.2015040102

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

Meged, A., & Gelbard, R. (2012). A unified fuzzy data model: Representation and processing. *Journal of Database Management*, *23*(1), 78–102. doi:10.4018/jdm.2012010104

Mironica, I., Ionescu, B., & Vertan, C. (2012). Hierarchical clustering relevance feedback for content-based image retrieval. *Proceedings of the10th International Workshop on Content-Based Multimedia Indexing (CBMI '12)* (pp. 1–6). doi:10.1109/CBMI.2012.6269811

Novak, D., Batko, M., & Zezula, P. (2011). Metric index: An efficient and scalable solution for precise and approximate similarity search. *Information Systems*, *36*(4), 721–733. doi:10.1016/j.is.2010.10.002

Tran, T., Phung, D., & Venkatesh, S. (2013). Learning sparse latent representation and distance metric for image retrieval. *Proceedings of theIEEE International Conference on Multimedia and Expo (ICME '13)* (pp. 1–6). IEEE.

Wang, L., Yang, L., & Tian, X. (2009). Query aware visual similarity propagation for image search reranking. *Proceedings of the17th International Conference on Multimedia (ACM Multimedia '09)* (pp. 725–728). doi:10.1145/1631272.1631398

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

Zikopoulos, P., & Eaton, C. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Education.

Zitouni, H., Sevil, S. G., Ozkan, D., & Duygulu, P. (2008). Re-ranking of web image search results using a graph algorithm. *Proceedings of the19th International Conference on Pattern Recognition (ICPR '08)* (pp. 1–4). IEEE. doi:10.1109/ICPR.2008.4761472

Petra Budikova is a post-doctoral researcher at the Faculty of Informatics, Masaryk University, Brno, Czech Republic, where she obtained a PhD Degree in computer science in 2013. In her research, she focuses on multimodal analysis of multimedia data with the utilization of content-based retrieval techniques. She is mainly interested in the synergy between the image and text modalities, and their utilization for image retrieval and annotation.

Michal Batko is an Assistant Professor at the Faculty of Informatics, Masaryk University, Brno, Czech Republic, where he obtained a PhD Degree in computer science. His research activities concentrate on the efficient searching in large distributed environments with emphasis on the scalability problem. The focus is especially on the problems of distributed similarity search using the metric space approach. As a software developer, he coordinates the development of an extensive similarity searching framework used for creating prototypes of indexing techniques and multimedia retrieval systems.

David Novak is a post-doctoral researcher at the Faculty of Informatics, Masaryk University, Brno, Czech Republic. He received a PhD in computer science from the same university and has been working on several national and European research projects. He is author of over 40 journal and conference papers mostly focusing on similarity-based indexing and searching, peer-to-peer and other distributed systems, and content-based multimedia retrieval. He is also interested in the areas of machine learning, information retrieval, and NoSQL databases. He spent the Fall semester 2015 as a Fulbright scholar with the CIIR group at UMass, Amherst, MA.

Pavel Zezula is a professor of informatics at the Faculty of Informatics, Masaryk University, Brno, Czech Republic. His professional interests concentrate on storage structures and algorithms for scalable content-based retrieval in non-traditional digital data types and formats. He has participated in numerous EU projects. His research team at the Masaryk University developed an extensible, scalable, and infrastructure independent similarity search engine. He is a co-author of more than 100 conference and 30 journal papers as well as the book Similarity Search: the Metric Space Approach, published by Springer US, 2006.