Fusion Strategies for Large-Scale Multi-Modal Image Retrieval

Petra Budikova, Michal Batko, and Pavel Zezula

Masaryk University, Brno, Czech Republic

Abstract. Large-scale data management and retrieval in complex domains such as images, videos, or biometrical data remains one of the most important and challenging information processing tasks. Even after two decades of intensive research, many questions still remain to be answered before working tools become available for everyday use. In this work, we focus on the practical applicability of different multi-modal retrieval techniques. Multi-modal searching, which combines several complementary views on complex data objects, follows the human thinking process and represents a very promising retrieval paradigm. However, a rapid development of modality fusion techniques in several diverse directions and a lack of comparisons between individual approaches have resulted in a confusing situation when the applicability of individual solutions is unclear. Aiming at improving the research community's comprehension of this topic, we analyze and systematically categorize existing multi-modal search techniques, identify their strengths, and describe selected representatives. In the second part of the paper, we focus on the specific problem of large-scale multi-modal image retrieval on the web. We analyze the requirements of such task, implement several applicable fusion methods, and experimentally evaluate their performance in terms of both efficiency and effectiveness. The extensive experiments provide a unique comparison of diverse approaches to modality fusion in equal settings on two large real-world datasets.

1 Introduction

Efficient management of multimedia data is quickly becoming a necessity in the current era of digital cameras, smart phones, and many other devices that allow people to produce and store enormous amounts of complex digital data. On one hand, the volumes of data currently available and the speed of its growth offer unprecedented resources for information mining and AI tasks. On the other hand, they also call for novel approaches to data organization that would be capable of dealing with large amounts of complex, heterogeneous content.

Although a number of multimedia search systems, both academic and commercial, have been created in recent years, the problem of effective and efficient retrieval still remains unsolved for many applications. Apart from the overall difficulty of the task, the multimedia retrieval field has also long suffered from a lack of suitable evaluation platforms and experimental data which made it difficult for researchers to analyze the strengths and weaknesses of individual solutions, especially in the context of large-scale retrieval. Recently, several large multimedia datasets have been made available for research purposes; however, the organization and evaluation of realistic benchmarking tasks still remains a demanding process and the existing comparisons only cover a limited scope of problems and techniques.

This work presents a comparative analysis of a set of state-of-the-art data management techniques that employ the multi-modal retrieval paradigm. Specifically, we focus on approaches that can be used in interactive large-scale multimedia searching. The selected class of methods is first studied on a theoretical level and then examined experimentally. The experimental evaluation is targeted on image data processing, but the principles described in the theoretical sections also apply to many other domains.

1.1 Evolution of Image Retrieval

Historically, the evolution of complex data retrieval can be traced from early attribute- and text-based solutions [6] to more recent content-based search strategies [26, 51] and finally the latest trends that combine multiple complementary techniques to obtain even better retrieval results [5]. The earlier approaches exhibit some very strong features but also significant drawbacks that prevented them from becoming a universal solution for complex data management. In particular, attribute-based and text-based searching can profit from mature database and text-retrieval technologies, but their usability is limited by the availability of descriptive metadata associated with the complex data objects. Content-based searching, on the other hand, exploits salient features that can be automatically extracted from data objects (e.g. a color histogram descriptor in case of image data). These are subjected to a suitable function that evaluates the similarity between pairs of objects, and the objects most similar to a given reference object are returned as the query result. A major advantage of this approach is the fact that no manually created metadata are necessary for supporting the data management; however, the content-based processing is often costly and approximate search strategies need to be applied to achieve online response times. Moreover, content-based searching often suffers from the *semantic gap* problem, i.e. the lack of correspondence between the information contained in the automatically extracted features and the human-perceived semantics of objects [83].

The most recent paradigm, termed *multi-modal (image) retrieval*, tries to overcome the limitations of the previous solutions by combining multiple complementary views on object relevance. This approach is very natural, as it follows the principles of human cognition processes [98]. Multi-modal retrieval techniques attempt to exploit as many information sources as possible, combining various content descriptors (e.g. color or shape descriptors in case of images) with context information available in different automatically captured metadata (e.g. EXIF, GPS location), text annotations, discussions on social networks, etc. [3, 22, 31, 41]. The multi-modal approach promises to improve the performance of retrieval systems on two levels: first, the limitations of any given modality should be reduced in the confrontation with other viewpoints on a candidate object's relevance; second, a well-designed multi-modal system should allow a complex evaluation of objects' relevance with acceptable costs, exploiting parallel processing of individual modalities and advanced filtering for fast and precise identification of candidate objects. To meet these expectations with a working search system, two principal questions need to be answered: 1) which data sources to select for a given application and how to extract maximum relevant information from them, and 2) how to combine these pieces of information effectively and efficiently.

Both these issues have been studied intensively in recent years and many solutions have been proposed for different use scenarios. However, a lot of work still remains to be done before the principles of multi-modal retrieval are sufficiently understood. One of the open problems concerns the practical applicability of different techniques that have been proposed for combining multiple modalities in the retrieval process. Although several studies that compare multiple techniques in equal settings have been published [11, 2, 30, 47, 63], none of them provides a systematic comparison of different multi-modal retrieval methods in the context of large-scale retrieval. Yet, the scalability aspect is extremely important in the Big Data era.

1.2 Our Contributions

Reflecting on this situation, the objective of this paper is to provide a systematic overview of existing multi-modal retrieval methods and analyze their properties with a special attention to their applicability for interactive large-scale searching. In the second part of the paper, we implement and experimentally evaluate selected fusion techniques over two large image datasets. The main contributions of the paper are the following:

- *Formal model of multi-modal retrieval*: We formally define the concept of modality and present a theoretical model of both mono-modal and multi-modal similarity-based retrieval, thus providing a solid foundation for our discussion of individual retrieval techniques.
- *Extended categorization of approaches to multi-modal retrieval*: Existing studies of multi-modal data management have established two basic categories of modality fusion techniques – the *early fusion* and the *late fusion*. Having analyzed a number of recent research works, we identify several additional aspects that are relevant for the practical applicability of multi-modal retrieval.
- *Analysis of modality fusion options for large-scale image retrieval*: Focusing on the specific task of large-scale image retrieval, we analyze its requirements and identify eligible fusion techniques.
- *Experimental evaluation of diverse approaches to large-scale image search*: Using a general framework for similarity-based data management and two large sets of real-world image data, we implement the selected techniques

and perform extensive experiments that allow us to compare these methods in terms of retrieval precision as well as processing costs. Using the experimental data, we derive some interesting insights that can be used for future optimization of multi-modal search systems.

The rest of this paper is organized as follows. Section 2 provides a formal background for the discussion of multi-modal retrieval techniques, which is followed by a survey and categorization of existing techniques in Section 3. In Section 4, we focus on the specific task of web-like image retrieval, discuss its requirements, and identify applicable techniques. Section 5 introduces our experimental framework and describes the implementation of selected methods. Section 6 details the evaluation procedure and experimental settings, the evaluation results are then reported in Section 7. Finally, Section 8 summarizes our findings.

2 Formal Model

Before we start analyzing possible approaches to modality fusion, let us formalize the basic concepts and processes that take part in the multi-modal retrieval. In this section, we first briefly review the basics of similarity-based searching model, which is a suitable abstraction of the retrieval process that covers all search systems in practical use. Next, we define a mono-modal retrieval model and then extend it to embrace multiple modalities.

2.1 Similarity Search

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 [98, 99]. 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 (so-called *trivial distance function*).

Let $\mathcal{X}$ be a collection of objects to be organized and $\mathcal{D}_{\mathcal{X}}$ be the domain of objects from $\mathcal{X}$. The similarity-based data retrieval exploits the "query-by-example" principle, where the query is defined by one or several reference objects $q_1, \ldots, q_n \in \mathcal{D}_{\mathcal{X}}$ and a similarity condition that needs to be satisfied by qualifying objects from $\mathcal{X}$. In this text, 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 a single reference point q. Nearest neighbor queries appear in many information retrieval tasks; apart from text search, kNN queries can be used to recognize a song from a fragment recording [35], track objects in videos [56], automatically cluster and annotate images [100], etc. Developing efficient and effective algorithms for kNN queries is thus a very important issue.

2.2 Single Modality Data Management

Let us now examine more closely the dataset $\mathcal{X}$, which contains objects of some generic data type $\mathcal{D}_{\mathcal{X}}$ (e.g. a vector, binary image, music recording, etc.). In many cases, objects from $\mathcal{D}_{\mathcal{X}}$ are very complex and not sufficiently structured to allow meaningful similarity evaluations. Images in particular can be seen as structured for storage and display, but are totally unstructured according to semantic content. Therefore, some suitable aspect of $\mathcal{D}_{\mathcal{X}}$ needs to be identified and used to represent each object for the purpose of data organization. We call each such aspect a *modality*, thus naturally extending the meaning of this term that originally referred to a physical representation of some information (e.g. text, video and sound capturing the same event). In the context of image retrieval, a typical example of a modality is a color histogram [62].

A modality $\mathcal{M}$ can be formally represented by an ordered pair $(p_{\mathcal{M}}, d_{\mathcal{M}})$ of a *projection function* $p_{\mathcal{M}} : \mathcal{D}_{\mathcal{X}} \rightarrow \mathcal{D}_{\mathcal{M}}$, where $\mathcal{D}_{\mathcal{M}}$ is a domain of modality $\mathcal{M}$, and a *distance function* $d_{\mathcal{M}} : \mathcal{D}_{\mathcal{M}} \times \mathcal{D}_{\mathcal{M}} \rightarrow \mathbb{R}_0^+$. The projection function can be applied on any object $o \in \mathcal{X}$ to extract a *feature descriptor* $o.f_{\mathcal{M}} \in \mathcal{D}_{\mathcal{M}}$, while the function $d_{\mathcal{M}}$ evaluates the distance between any two descriptors, i.e. the dissimilarity of the respective objects as seen in the view of modality $\mathcal{M}$.

Let $SE_{\mathcal{M}}$ be a mono-modal search engine that uses a single modality $\mathcal{M}$ for data organization. Typically, $SE_{\mathcal{M}}$ stores each object $o \in \mathcal{X}$ as a pair $(o.f_{\mathcal{M}}, o)$ and uses $o.f_{\mathcal{M}} = p_{\mathcal{M}}(o)$ to search for the data object o. The search engine $SE_{\mathcal{M}}$ may employ one or several index structures $I^1_{\mathcal{M}}, \dots, I^n_{\mathcal{M}}$ that organize the descriptors of objects in $\mathcal{X}$ [12, 79, 99]. A similarity query over $SE_{\mathcal{M}}$ is defined by a query object $q_{\mathcal{M}}$, which needs to be from the domain $\mathcal{D}_{\mathcal{M}}$ (clearly, such query can be easily extracted from a more user-friendly query object $q_{\mathcal{X}} \in \mathcal{D}_{\mathcal{X}}$, as depicted in Figure 1). The kNN query is then defined as follows:

$$kNN_{\mathcal{M}}(q_{\mathcal{M}}, \mathcal{X}) = \{\mathcal{R} \subseteq \mathcal{X}, |\mathcal{R}| = k \wedge \forall x \in \mathcal{R}, y \in \mathcal{X} \setminus \mathcal{R} : d_{\mathcal{M}}(q_{\mathcal{M}}, p_{\mathcal{M}}(x)) \leq d_{\mathcal{M}}(q_{\mathcal{M}}, p_{\mathcal{M}}(y))\}$$

2.3 Multi-Modal Data Management

Although different sophisticated modalities have been proposed for images and other types of complex data, experience shows that each modality has some limitations that prevent it from fully answering to users' needs [5]. Some modalities do not sufficiently capture the user-perceived similarity of the original objects (e.g. the color histogram), other are highly computationally demanding (e.g. various local visual features) or not available in all situations (e.g. descriptive text). To overcome this problem, multi-modal data management systems employ a set of modalities $\mathcal{M}_1, \dots, \mathcal{M}_n$ that are relevant for the given data domain $\mathcal{D}_{\mathcal{X}}$ and the target applications. The modalities can be combined in many different ways to provide more complex representations of objects from $\mathcal{X}$ and to evaluate their similarity on a higher semantic level.

To describe the functionality of a multi-modal system, we need to introduce additional notation. Let $p_{\widehat{\mathcal{M}_{i_1}, \dots, \mathcal{M}_{i_m}}}$ be a multi-modal projection function that

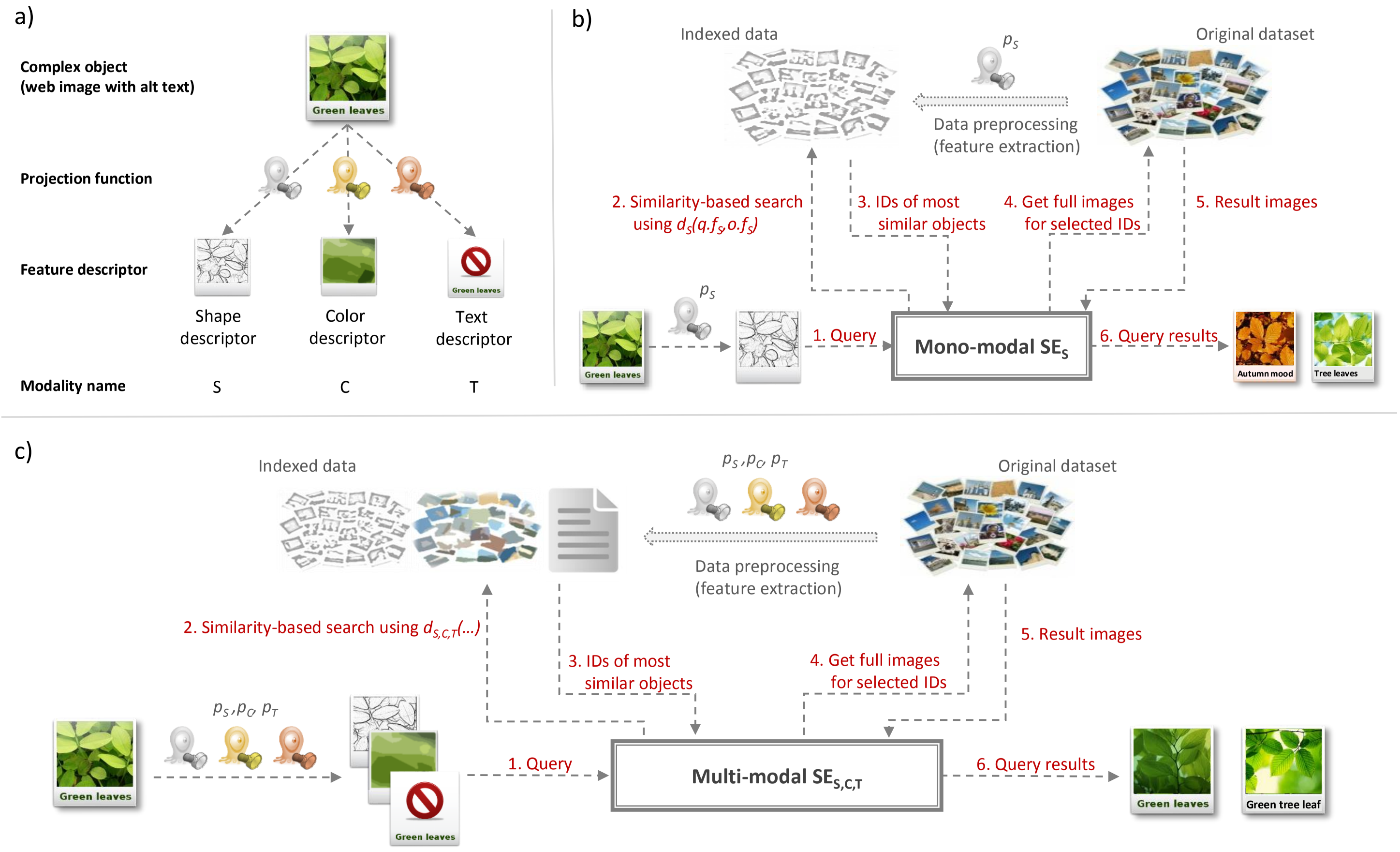

Fig. 1. Web image search example: a) Modalities, b) single-modality retrieval, c) multi-modal retrieval.

transforms an object $o \in \mathcal{X}$ into a multi-modal descriptor $o.f_{\widehat{\mathcal{M}_{i_1},\ldots,\mathcal{M}_{i_m}}}$; trivially, $o.f_{\widehat{\mathcal{M}_1,\mathcal{M}_2}}$ can be obtained by concatenation of $o.f_{\mathcal{M}_1}$ and $o.f_{\mathcal{M}_2}$, but more sophisticated techniques are also available. Similarly, let $d_{\widehat{\mathcal{M}_{i_1},\ldots,\mathcal{M}_{i_m}}}$ be a multi-modal distance function that evaluates the dissimilarity of objects with respect to multiple viewpoints; again, $d_{\widehat{\mathcal{M}_{i_1},\ldots,\mathcal{M}_{i_m}}}$ can be defined in many ways, which will be discussed later.

Now, we are able to define a multi-modal search engine $SE_{\mathcal{M}_1,\ldots,\mathcal{M}_n}$ that recognizes a set of modalities $\mathcal{M}_1,\ldots,\mathcal{M}_n$. $SE_{\mathcal{M}_1,\ldots,\mathcal{M}_n}$ is characterized by a set of projection functions π and a set of distance functions δ that can be used to organize objects from $\mathcal{X}$. Set π contains all supported projection functions over $\{\mathcal{M}_1,\ldots,\mathcal{M}_n\}$; similarly, δ contains all supported distance functions. $SE_{\mathcal{M}_1,\ldots,\mathcal{M}_n}$ may further exploit a set of multi-modal index structures $\iota_{\widehat{\mathcal{M}_{i_1},\ldots,\mathcal{M}_{i_m}}}$, which can be used to retrieve candidate objects relevant with respect to the particular modalities engaged.

A query $Q = (q, d_Q)$ over $SE_{\mathcal{M}_1,\ldots,\mathcal{M}_n}$ is defined by a query object q and a distance function d_Q. The query object q can be specified as $q_{\mathcal{X}} \in \mathcal{D}_{\mathcal{X}}$, by a single modality descriptor $q_{\mathcal{M}_i} \in \mathcal{D}_{\mathcal{M}_i}$, or as a combination of several modality descriptors $(q_{\mathcal{M}_{i1}},\ldots,q_{\mathcal{M}_{im}})$. The query distance d_Q needs to be taken from the set δ of supported distance functions.

3 Categorization of Approaches

Having defined the multi-modal search paradigm, we can now proceed with a more detailed study of different projection and distance functions and the specific techniques of modality fusion. At the same time, we introduce a new categorization of existing multi-modal search methods in this section. Some of the observations presented here have been inspired by discussions of fusion techniques in multimedia processing survey studies [5, 14, 28, 47] and also by several research works that deal with information fusion in different domains [13, 75, 76]. However, to the best of our knowledge no other taxonomy of fusion methods exists that would take into account all the factors discussed here.

Our categorization is defined by several dimensions of the fusion that we believe to be significant for large-scale retrieval. These dimensions are not orthogonal but rather interconnected, so that a single design decision often influences several of the properties we study. However, we prefer to analyze the individual aspects separately to see more clearly how the different types of solutions work and what are their strengths and weaknesses. The dependencies between individual dimensions will naturally be mentioned in the discussions and summarized at the end of this section.

3.1 Integration of Modalities

The fundamental idea of multi-modal search paradigm is to exploit several complementary modalities $\mathcal{M}_1,\ldots,\mathcal{M}_n$ to describe complex data objects and evaluate their similarity. During data processing and query evaluation, these modalities need to be combined together to produce the overall similarity measure d_Q requested for a particular query Q. The fusion process may take into account the individual data descriptors $f_{\mathcal{M}_1}, \ldots, f_{\mathcal{M}_n}$, the respective distance functions $d_{\mathcal{M}_1}, \ldots, d_{\mathcal{M}_n}$, or both. In this section, we focus on the semantics of different approaches to modality integration.

Among the existing solutions, we can distinguish two classes of methods that differ significantly with regard to the relative importance assigned to individual modalities during the retrieval process. In the *symmetric fusion* paradigm, all modalities are considered to be equally important for the data management and are utilized in all phases of query processing. In *asymmetric fusion*, some of the modalities are treated as more influential and are used to organize and pre-select data, while the remaining modalities are only used for query result refinement. The choice between these two options, and the subsequent selection of integration parameters, depends on various properties of the input modalities, the target application characteristics, and efficiency requirements.

Symmetric Fusion In solutions that follow the symmetric fusion paradigm, all modalities are considered independent and can be processed in parallel until the moment of fusion, when all of them are merged together. Even though the contribution of each modality can be increased or decreased by a particular setting of the fusion mechanism, it is important that all modalities are used for indexing and searching of the whole dataset $\mathcal{X}$. The following sections present possible implementations of this fusion type.

Feature fusion Feature (or descriptor) fusion is an integral part of early fusion strategies, which combine modalities $\mathcal{M}_1, \ldots, \mathcal{M}_n$ prior to data indexing. The joining of modalities is applied on the level of descriptors, where individual mono-modal descriptors $o.f_{\mathcal{M}_1}, \ldots, o.f_{\mathcal{M}_n}$ of a given data object are merged into a single complex descriptor $o.f_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}}$. This descriptor is provided by a suitable multi-modal projection function $p^{FF}_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}} : \mathcal{D}_{\mathcal{X}} \rightarrow \mathcal{D}_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}}$, and the similarity of two objects is evaluated by a multi-modal distance function $d^{FF}_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}} : \mathcal{D}_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}} \times \mathcal{D}_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}} \rightarrow \mathbb{R}^+_0$.

For a simple feature fusion, individual mono-modal descriptors can be straightforwardly concatenated to form the multi-modal descriptor. In the first multimodal solutions for ImageCLEF retrieval tasks [28] or video retrieval [84], the concatenated descriptors were perceived simply as points of a multi-dimensional vector space and standard L_p metrics were applied to measure their distance. However, this approach may degrade the performance of multimedia content analysis algorithms, especially when the features are independent or heterogenous [84]. Therefore, most systems that employ feature concatenation combine it with the distance aggregation approach that will be discussed in the following section.

If training data is available for a given retrieval task, it is possible to engage more advanced feature fusion strategies. These define p^{FF} by mining semantic relationships between modalities and identification of data characteristics that are most important with respect to a given data set and/or retrieval task [31, 32, 36, 57, 72, 71, 82, 90, 92]. As detailed in [5], the most common sematic fusion methods include SVMs, Bayesian models, neural networks. The resulting feature space typically has a lower number of dimensions than the input ones, therefore the feature fusion also serves as a dimensionality reduction technique. A suitable distance function can also be determined by the semantic analysis [94].

Distance aggregation Alternatively, it is possible to perform symmetric fusion by combining partial distances of object $o \in \mathcal{X}$ measured by $d_{\mathcal{M}_1}, \ldots, d_{\mathcal{M}_n}$, using a suitable aggregation function $d^{\underline{AGG}}_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}} : (\mathbb{R}_0^+)^n \to \mathbb{R}_0^+$. The aggregated distance can be combined with previously described feature concatenation in early fusion systems, however its main use is in late fusion architectures where each modality is indexed separately. As will be discussed later, some properties of the aggregation function (e.g. monotonicity) may be crucial for the selection of the late fusion method. In the late fusion phase, it is also possible to combine object ranks from previous retrieval phases instead of the actual distances. Recent study [70] suggests to use both the ranks and distances during the fusion.

Similar to feature fusion, there exist several categories of distance aggregation methods. The first category is composed of so-called *blind* fusion functions [67], where fixed rules are applied regardless of individual distance function value distributions. Examples of such aggregations include non-weighted min, max, sum, product, or geometric mean [28, 67]. The second class contains weighted linear and non-linear aggregation functions, the most popular of which is definitely the weighted sum of $d_{\mathcal{M}_1}(p_{\mathcal{M}_1}(o), p_{\mathcal{M}_1}(q)), \ldots, d_{\mathcal{M}_n}(p_{\mathcal{M}_n}(o), p_{\mathcal{M}_n}(q))$ [8, 55, 93]. Other possible options include weighted product, sum of logarithms, sum of squares, sum of k lowest distances, etc. [5, 24, 54, 102]. Aggregation parameters such as the weights of individual modalities can be determined by domain experts or dataset analysis and machine learning [8, 93]. Alternatively, users can personalize the search by setting the respective weights manually, if the system architecture supports flexible aggregations (see Section 3.3). In [49, 102], the authors attempt to determine optimal fusion coefficients dynamically for individual queries without user interaction. Also, it is often necessary to normalize the individual distances before the aggregation, which is studied in [7]. Finally, the most complex aggregation functions engage probabilistic or regression models of distance distributions [5].

Asymmetric Fusion Asymmetric fusion strategies constitute a complement to the symmetric solutions. Here, the modalities $\mathcal{M}_1, \ldots, \mathcal{M}_n$ are not considered equal but instead, one or several of the modalities are chosen as dominating or *primary*. Let us suppose that the modalities are ordered in such a way that $\mathcal{M}_1^P, \ldots, \mathcal{M}_m^P$ are the primary ones. These modalities are used in data indexing phase to organize dataset $\mathcal{X}$, and in a search session to pre-select a set of candidate objects $CS_{\mathcal{M}_1^P,\ldots,\mathcal{M}_m^P}$. This candidate set is then subjected to further evaluation, where *secondary* modalities $\mathcal{M}_{m+1}^S, \ldots, \mathcal{M}_n^S$ as well as the primary ones may be exploited. Noticeably, such solution typically results in an approximate retrieval, where the query result $\mathcal{R}$ is evaluated as follows:

$$\mathcal{R} = kNN_{\mathcal{M}_1,\ldots,\mathcal{M}_n}(Q, CS_{\mathcal{M}_1^P,\ldots,\mathcal{M}_m^P}), \text{ where}$$
$$CS_{\mathcal{M}_1^P,\ldots,\mathcal{M}_m^P} = \kappa NN_{\mathcal{M}_1^P,\ldots,\mathcal{M}_m^P}(Q', \mathcal{X})$$

Here, the κ denotes the size of the candidate set $CS_{\mathcal{M}_1^P,\ldots,\mathcal{M}_m^P}$ and Q' the query object transformed into the domains of values of the primary modalities. For obvious reasons, this approach is also denoted as incremental data *filtering*. Parameter κ significantly influences both the evaluation costs and the precision of results, therefore its value needs to be chosen carefully [4].

The motivation for applying the asymmetric fusion may be threefold: 1) the primary modalities are more important for the user – this is typical e.g. for location-aware applications; 2) the asymmetric solution is chosen because of efficiency issues – e.g. text search is a very efficient method that is often used to pre-select the candidate set for further processing; or 3) some of the modalities may not be available at the beginning of the query evaluation but emerge later by means of (pseudo)-relevance feedback. In the first two situations, the primary and secondary modalities can be fused in any of the ways mentioned above. A typical asymmetric fusion system is composed of a text-based primary search (possibly over several text features joined by feature fusion) and a re-ranking phase, during which distance aggregation over several modalities is applied [11, 28]. In addition to this, most of the recent asymmetric fusion solutions exploit the pseudo-relevance feedback principle, which allows to introduce *context-aware* modalities in the second retrieval phase. These are defined by projection and distance functions that take into account the properties of the actual candidate set $CS_{\mathcal{M}_1^P,\ldots,\mathcal{M}_m^P}$ and the relationships between objects in this set. The idea of context-aware modalities is based on the assumption that objects relevant to a given query should be similar to each other, while the less relevant ones are likely to be outliers in a similarity graph of objects from $CS_{\mathcal{M}_1^P,\ldots,\mathcal{M}_m^P}$. This assumption is exploited by many clustering-based distance measures [39, 40, 61, 68, 103], distances based on random walks in the similarity graph [44, 45, 52, 64, 74, 89, 101], and several other contextual distance measures [43, 70]. More detailed discussion of re-ranking techniques can be found in surveys [2, 60].

3.2 Fusion Scenarios

By the term *modality fusion scenario*, we denote the sequence of actions that are undertaken by the system during data organization and query processing in order to combine the modalities. The fusion scenario is a principal characteristics of multi-modal systems that strongly influences the overall efficiency and effectiveness. Traditionally, multi-modal approaches are divided into two classes denoted as *early fusion* and *late fusion*. In this section, we take a closer look on individual solutions within each of these classes, and define a more fine-grained categorization of late fusion methods.

Early Fusion: Data Preparation and Indexing Under the early fusion paradigm, modalities $\mathcal{M}_1, \ldots, \mathcal{M}_n$ are combined prior to data indexing. After initial data analysis, the search system employs a single fused projection function $p_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}}$ and a distance measure $d_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}}$, which can be understood as a new fused modality. Early fusion is also denoted as *data fusion*, *feature fusion*, or a *joint features model*, because it happens on the feature level, before any decisions concerning the similarity of objects are taken. The early fusion is in principle a symmetric approach. Any of the fusion techniques surveyed in Section 3.1 can be employed to provide $p_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}}$ and $d_{\widehat{\mathcal{M}_1,\ldots,\mathcal{M}_n}}$, the best results are naturally achieved when semantic analysis of relationships between modalities is used [28, 31, 32, 36, 87].

A great strength of the early fusion paradigm is the fact that the fusion process can exploit rich information about the whole dataset $\mathcal{X}$ (in contrast to late fusion methods which typically work with pre-filtered data), and the modality fusion is performed off-line. This allows to thoroughly analyze the data, construct optimal fused projection and distance functions, and build an optimal index structure for the new modality. On the other hand, a major disadvantage of early fusion solutions is the limited flexibility of the resulting search system. The combination of modalities is usually fixed in the index and cannot be adjusted to accommodate particular user's preferences. Even though some progress has been made towards providing index structures that support multiple distance functions [18, 23], the flexibility is still very limited. Moreover, sophisticated early fusion methods that analyze semantic relationships between modalities require high-quality training data, which is often difficult to obtain, and substantial computational resources, which may become a limitation of scalability.

Late Fusion: Query Evaluation In a multi-modal search system that exploits late fusion, modalities $\mathcal{M}_1, \ldots, \mathcal{M}_n$ are not fused in advance, but only during query evaluation. This approach can be perceived as an on-request fusion – a late fusion system typically supports mono-modal retrieval over some of the available modalities as well as different settings of multi-modal searching. The resulting flexibility of searching is one of the most important benefits of late fusion.

In modern retrieval systems, the query evaluation is often a complex and possibly iterative process. As depicted in Figure 2, there are several common query processing phases that differ in the amount and type of information that is exploited there. In the following sections, we briefly describe each of these phases and discuss modality fusion techniques that can be implemented in individual phases to refine the query $Q = (q, d_Q)$ and to identify relevant objects from $\mathcal{X}$.

Query specification and preprocessing In the beginning of the retrieval process, users need to express their information need as a query. This is composed of a (multi-modal) query object q (i.e. an example image and a set of keywords) and a distance function d_Q to be used for selection of similar objects. Before the query is submitted to the search system, different preprocessing techniques may be applied to refine, disambiguate, or expand the query [6, 19].

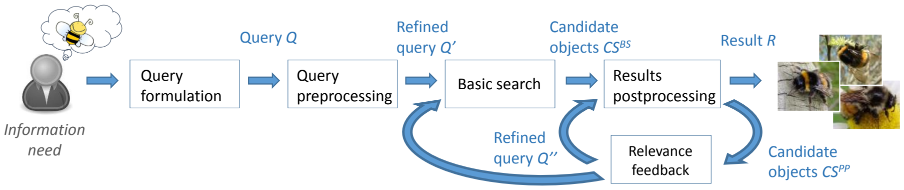

Fig. 2. Phases of query evaluation.

In the context of multi-modal query preprocessing, the synergy between modalities can be exploited for both query disambiguation and expansion. Typically, the preprocessing introduces an auxiliary query, which may be evaluated over the target dataset or some external knowledge base (e.g. WordNet [34] or ImageNet [27] for image-and-text query preprocessing). The disambiguation process is used to refine mono-modal descriptors provided by users and thus replace the query object q by q' [1, 58], whereas query expansion retrieves additional modalities from external resources, producing a more complex query object q' that requires a new distance function d'_Q [22, 86].

Basic search A fundamental part of any query evaluation is the *primary* or *basic search (BS)*, during which a candidate result set CS^{BS} is selected from the whole dataset $\mathcal{X}$. Depending on the strategy of the search engine, CS^{BS} may be either directly presented as the final result, or submitted to a postprocessing phase. In the latter case, CS^{BS} is usually several orders of magnitude larger than the requested result set.

In late fusion systems, the dataset $\mathcal{X}$ is typically preprocessed (indexed) using one or several separate modalities. Let independent indexes $I_{\mathcal{M}_{i1}}, \ldots, I_{\mathcal{M}_{im}}$ be available in $SE_{\mathcal{M}_1,\ldots,\mathcal{M}_n}$. During the basic search phase, some of these can be utilized for standard mono-modal retrieval and produce intermediate results on which the fusion will be applied in a latter processing phase. Alternatively, modality fusion can be implemented during the basic search phase, both in a symmetric and asymmetric manner.

The symmetric basic-search-phase fusion is best represented by the Threshold Algorithm [33], which works as follows. For each modality $\mathcal{M}$ to be fused, there needs to exist an index from which individual objects $o \in \mathcal{X}$ can be retrieved one by one, ordered by their increasing distance $d_{\mathcal{M}}(p_{\mathcal{M}}(q), p_{\mathcal{M}}(o))$ from the query object in the view of modality $\mathcal{M}$. Apart from this *sorted access* to objects, there also has to be a *random access* method that can access any object from $\mathcal{X}$ and retrieve its aggregated distance from the query, taking all modalities in consideration. The Threshold Algorithm then proceeds in iterations: in each iteration, the next object is retrieved from each index and the aggregated multi-modal query distance d_Q^{AGG} is computed for each retrieved object. After each iteration, the intermediate query result is updated to contain the best objects seen so far, and a stopping condition is evaluated that decides whether better results can be found among the yet unseen objects. The stopping conditions compares the aggregated distance of the most dissimilar object in the intermediate result with the *threshold value* (lower bound) on the distance of yet unseen objects, which is computed by applying the aggregation function on the highest partial distances seen so far in each of the mono-modal sorted lists. Provided that the aggregated distance function d_Q^{AGG} is monotonous, the algorithm guarantees that the most relevant objects from $\mathcal{X}$ with respect to d_Q^{AGG} are found. However, this method can run into performance problems since the number of objects from $\mathcal{X}$ that may need to be accessed is not known in advance. Therefore, approximate implementations have also been studied [9, 33].

For asymmetric basic-search-phase fusion, it is necessary to utilize a specialized index structure constructed in such way that the data is organized by one modality but can be searched by a combination of several. This can be achieved by extending a standard mono-modal index with additional information about secondary modalities and adjusting the retrieval algorithm so that it takes these modalities into account when pruning the search space and identifying candidate objects. Solutions of this type have been studied mainly in the context of geo-textual data processing. For instance, the IR-tree [25] 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. Any monotone aggregation function can be then used to compute the query distance. Several other geo-textual indexes are analyzed in [21], a more generic but approximate solution with metric-based index was proposed in [15].

Result postprocessing When the postprocessing phase is implemented, its task is to re-evaluate the similarity between the query and the objects in CS^{BS}, using more complex measures of similarity. No more objects are accessed in the postprocessing phase than those in CS^{BS}, therefore the postprocessing is often denoted as *result ranking* or *re-ranking.*

According to [28], postprocessing-phase fusion is the most frequent type of multi-modal image search, and the same trend can be observed in other multimedia retrieval fields. The main advantages of this solution are its low processing costs, the possibility to apply fusion on top of well-established mono-modal index structures, and no limitations on the form of the aggregated query distance. Postprocessing fusion is also very often combined with pseudo-relevance feedback, which allows to exploit context-aware modalities.

When CS^{BS} is provided by multiple indexes/modalities, the postprocessing fusion is symmetric and approximates the Threshold Algorithm. Instead of accessing all potentially relevant objects from $\mathcal{X}$, only a fixed number of top-ranking objects is retrieved from each index, merged, and re-ranked. Examples of symmetric postprocessing fusion include text-and-visual fusion in [28], combination of different visual modalities [20], or a fusion of multiple text-based search results [54].

However, the more typical type of postprocessing fusion is asymmetric combination of modalities. One modality with low processing costs and good selectivity is chosen for indexing and basic search, while the remaining ones are utilized during re-ranking. Most commonly, text is used as the primary modality [4, 28, 44, 64, 81, 89], but some solutions that utilize content-based retrieval as the primary modality also exist [17, 53, 65, 80]. A more detailed survey of re-ranking mechanisms and comparison of selected techniques can be found in [11, 2, 60].

Relevance feedback Relevance feedback is a result refinement mechanism that assumes interactive searching, where users repeatedly provide their opinion on the relevance of candidate objects [77]. In *pseudo-relevance feedback* variant, user opinion is replaced by assumption that candidate objects from the last iteration are likely to be relevant and their properties can be used to predict the properties of the desired answer. With both interactive and automatic evaluation, the feedback loop may be repeated several times. In each iteration, either the query object or the query distance measure is updated. The refined query is then reintroduced either to the basic search, or the result postprocessing phase.

In the context of modality fusion, relevance feedback may be utilized to obtain values of some modalities that are not present in the query specification, to refine the values of available modalities, or to adjust the query distance function to better suit the user's information need [53, 61, 85, 88, 95–97]. The most frequent pseudo-relevance feedback methods are based on candidate set clustering and random walks in a candidate objects' similarity graph, as discussed in Section 3.1.

Comparison of Early and Late Fusion While state-of-the-art research literature provides many examples of both early and late fusion methods, there are not many guidelines for deciding which approach is more suitable for a given application. The effects of early and late fusion on retrieval result quality have been compared in several studies, but the results are not very conclusive. Some authors find the early fusion to be superior since it allows complex semantic analysis of the data [30, 84], others conclude that late fusion can provide better results in many situations [24, 28]. In principle, early fusion is likely to provide good results if the user/application needs are well understood and good training data is available, which is satisfied e.g. for well-defined classification tasks [48]. On the other hand, late fusion should be preferred in general-purpose retrieval where users are expected to interact with the system and adjust the evaluation of similarity to their preferences. Late fusion is also a natural implementation for asymmetric integration of modalities. Finally, some authors propose to combine early and late fusion to achieve the best results [50].

3.3 Flexibility

As suggested in the previous section, one of the big challenges of searching in broad data domains is the fact that it is impossible to define a universal similarity measure that would be suitable across different queries and user needs. This introduces the need for flexible retrieval methods that would allow users to influence the choice of modalities and the manner in which they are combined.

As we have observed in the descriptions of the fusion scenarios, not all modality fusion techniques allow users to adjust the combination. Typically, early fusion approaches do not support flexible searching, whereas some late fusion architectures are highly adaptable. We propose to distinguish the following three levels of flexibility.

Zero flexibility In zero flexibility systems, the selection of modalities as well as their combination is fixed. This applies for most early fusion systems [3, 8, 32, 36, 72, 82, 90, 92].

Aggregation flexibility In this case, the selection of modalities is fixed, but users can influence the aggregation function. The aggregation flexibility can be either *full*, or *partial*. In the latter case, the set of supported aggregation functions is limited by some required properties (e.g. monotonicity is needed for [23, 33]). Full aggregation flexibility is provided by most postprocessing fusion solutions [4, 44, 89, 60].

Feature flexibility Again, we distinguish between a *full* and *partial* feature flexibility. For full flexibility, the modalities to be fused need not be known in advance, since users can introduce additional modalities during query specification. The system has to be able to embrace the new modalities without rebuilding the whole search infrastructure, which is easily achieved in postprocessing fusion. In case of a partial feature flexibility, adding a new modality needs to be processed off-line and may require adaptations of the infrastructure, but does not necessitate a complete rebuild of the search system. This is satisfied by asymmetric indexing structures such as the IR-tree [21, 25].

3.4 Precision

The precision of any search result can be analyzed from two different perspectives: 1) a *distance-based* or *objective* perspective analyses the result precision with respect to the selected data representation and the query distance function d_Q, whereas 2) a *user-perceived*, *subjective*, or *semantic* perspective takes into account the users' satisfaction with the result. The second view determines the real usability of the respective search system, but depends on multiple factors – the selection of modalities, quality of data capturing and feature extraction, definition of the distance function, and the objective precision of the actual retrieval – and can only be assessed by user-satisfaction studies. In this section, we focus only on the distance-based precision, which can be objectively measured.

The objective retrieval precision of 100 % can be always achieved by exhaustive checking of all objects in $\mathcal{X}$. However, in large-scale searching some approximations are usually applied during the query evaluation to decrease the computation costs. These approximations may not result in any noticeable deterioration of user-perceived result quality, as the similarity-based searching is (semantically) approximate by nature. Still, the distance-based approximation ratio should intuitively not be too large if we do not want to risk decreasing user satisfaction. In the multi-modal retrieval, we can identify two types of distance-based approximations: those that regard the processing of individual modalities, and approximations of the actual fusion that can be applied when late fusion strategies are used. Single-modality retrieval approximations are analyzed in a survey study [69], which identifies several important aspects of approximation strategies. In the following, we study the same aspects for fusion approximations.

Applicability of a given technique on different data domains The applicability of fusion solutions is very wide in case of postprocessing fusion (e.g. [4, 44, 89, 60]), whereas basic search fusion is more restricted. Specifically, basic search fusion mechanisms either pose limitations on the aggregation function (e.g. the Threshold Algorithm [33]) or are suitable only for specific data and distance function (e.g. the geo-textual indexes [21, 25]).

The principle of achieving approximation From the implementation point of view, most of the fusion approximations fall into the category of *reducing comparisons* – the similarity of objects is not evaluated for all candidates that are potentially relevant, but only for such objects that are considered most promising in a given processing phase.

Result quality guarantees Considering the quality of the results, existing fusion techniques either guarantee 100 % fusion precision (early fusion, the Threshold Algorithm) or give no guarantees on quality apart from reporting experimental results (postprocessing fusion techniques).

User interaction with the system The majority of approximate fusion techniques allow users to influence the trade-off between retrieval costs and precision, e.g. by setting the size of CS^{BS}.

3.5 Efficiency and Scalability

Retrieval efficiency and search system scalability are clearly crucial qualities of any system designed for big data processing. In multimedia retrieval, there are two major issues that need to be addressed: the costs of data preprocessing, and the efficiency of query evaluation. Depending on a selected fusion scenario, multi-modal retrieval introduces additional costs to one or both of these phases.

Data preprocessing In the data preprocessing step, descriptors of primary modalities first need to be extracted from all objects in $\mathcal{X}$. This complexity of the extraction process is linear with respect to the size of $\mathcal{X}$, with the actual costs depending on the selection of modalities – when sophisticated content-based descriptors are used, the extraction process can be very computationally intensive [8]. In early fusion scenarios, the extracted descriptors are immediately analyzed and fused. The costs of this phase depend on the specific fusion technique employed, but the complexity of semantic fusion techniques is in general super-linear with respect to the size of $\mathcal{X}$. Finally, index structures for either the original or the fused descriptors are created [12, 79, 99].

Even though the data preprocessing phase is evaluated off-line, its complexity may become a bottleneck of the overall system scalability. Efficient extraction of descriptors for very large data is considered a challenging task nowadays [98], so the choice of primary modalities should be made carefully. To the best of our knowledge, complex early fusion has never been implemented in very large scale.

Query evaluation with early fusion As discussed earlier, query evaluation in early fusion systems is equal to mono-modal query evaluation. After the extraction of query descriptors, which requires constant time, relevant objects from $\mathcal{X}$ are identified in the index. The complexity of index retrieval is usually strongly sublinear or even constant, depending on the level of approximation applied [66].

Query evaluation with late fusion In case of late fusion, additional processing is added to the query evaluation costs. The actual fusion complexity is determined by the number of objects that are considered during fusion. As discussed in Section 3.2, in basic search fusion the number of objects may be unlimited and the fusion may thus degrade to linear complexity with respect to the size of $\mathcal{X}$. On the other hand, the number of objects entering postprocessing fusion is always limited. The postprocessing costs may be high with respect to the size of CS^{BS}, but are constant with respect to the size of $\mathcal{X}$. Furthermore, the efficiency of late fusion is influenced by the choice between symmetric and asymmetric integration of modalities. Symmetric fusion may exploit parallel processing, whereas asymmetric fusion typically tries to minimize processing costs by utilizing cheap and highly selective modalities first.

3.6 Axis Correlations & Other Aspects

In the previous sections, we have defined five axes that can be used to classify multi-modal retrieval techniques. As already mentioned, these axes are not orthogonal; on the contrary, a single design decision typically determines several of the axes. The correlations between individual axes were discussed in the descriptions of individual axes. To complete our analysis, Table 1 presents an overview of meaningful combinations of individual approaches.

The five classification criteria that we have introduced represent important characteristics of large-scale multi-modal retrieval, but are by no means exhaustive. Many other aspects are worth attention and need to be considered carefully to design a multi-modal search system. The *selection of modalities* is extremely important – it is necessary to choose such a set of modalities that provides complementary information and does not require too costly processing [5, 63]. *Individual application domains* may require different modalities and pose various restrictions on evaluation costs, precision and flexibility. The *level of user participation* may also vary; in general, users do not like to provide much input during the query processing, but in some cases intensive interaction can be expected. *Additional information sources* such as ontologies or general web data need to

Fusion strategy	Flexibility	Approxim.	Scalab.	Examples
Symmetric early fusion				
Simple early fusion	zero	none	medium	[3, 8, 84]
Semantic early fusion	zero	none	medium or high	[31, 32, 36, 57, 71, 72, 82, 87, 90, 92, 94]
Multi-metric indexing	partial aggreg. f.	none	medium	[18, 23]
Symmetric late fusion				
Threshold algorithm (basic search phase)	partial aggreg. f. partial feature f.	none or guaranteed	low	[33]
Symmetric postprocessing	full aggreg. f. partial feature f.	not guaranteed	high	[5, 20, 24, 54, 67, 78]
Asymmetric late fusion				
Asymmetric indexing (basic search phase)	partial aggreg. f. partial feature f.	none	medium	[15, 21, 25]
Asymmetric postprocessing	full aggreg. f. full feature f.	not guaranteed	high	[4, 11, 17, 39, 40, 43–45, 52, 53, 61, 64, 65, 68, 70, 73, 74, 80, 81, 85, 88, 89, 95–97, 101–103]

Table 1. Multi-dimensional classification of fusion techniques.

be studied, collected or created, cleaned, and maintained [37, 38]. Finally, the *synchronization of modalities* is vital for modalities with a time dimension, e.g. sound or video [5].

4 Large-scale Multi-Modal Image Search

In the previous section, we have surveyed fundamental characteristics of multi-modal searching, taking into account a number of diverse methods that have been proposed for different situations. The presented categories provide us with basic guidelines that can be used to select possible solutions for a given application. However, the multi-modal retrieval is a complex task with many factors that influence the search results, thus the true usefulness of any method cannot be determined unless it is experimentally verified in context of the target use case and modalities. Each such evaluation also provides new data that allow the scientific community to study relationships between information needs, modalities, and retrieval techniques.

Accordingly, the second part of this paper is devoted to an experimental evaluation of techniques applicable for interactive large-scale image search, which is a principal component of many popular applications, e.g. web galleries, social networks, etc. Specifically, we focus on the fusion of text and visual modalities in this context. In a typical browsing scenario, a user sees an image for which he or she would like to get similar ones. The usual way to provide the results is to execute a textual search based on the annotation of the original image. However, since the user selected a particular image, the visual content of the image is also important. Thus a multi-modal search combining the textual and visual aspect search is likely to produce better results. Even though search engines with such functionality already exist, their design is often based on assumptions and expectations that have not been rigorously defined and studied. To address this situation, we perform an extensive evaluation of different approaches to image-and-text retrieval and analyze the results.

4.1 Review of Requirements

First, let us briefly analyze the basic characteristics of web image searching. Web search, as opposed to retrieval from specialized resources, is often used by people who do not know precisely what they are searching for. These users are looking for inspiration or some general information (i.e. "browsing") rather than performing a "targeted search" for a specific item [26, 46]. User's preferences tend to become more focused during a search session, when the results that are found influence the user [98].

The implications of this behavior are two-fold. On one hand, user's uncertainty about the desired result relaxes the requirements for objective precision – the results need to be relevant, but not necessarily the most relevant items that exist for a query that in itself is often just an approximate expression of user's information need. On the other hand, there are strong requirements concerning search efficiency and flexibility. Efficiency is crucial for user's convenience, especially when a search session consists of more than one query-response cycle. As for flexibility, we have already debated that it is impossible to define a universally applicable model of similarity for a broad-domain searching. Even though current search engines provide only limited means of adjusting the retrieval semantics, the flexibility of searching is becoming one of the most important features that are required e.g. for personalized searching.

4.2 Task Specification

For the purpose of experimental evaluation, we define the large-scale image retrieval task as follows. We assume a k-nearest-neighbor search, where the user issues a multi-modal query and expects k relevant images. We only consider a single iteration of a query session. To keep the evaluation task feasible, we only employ the two most popular image search modalities – the textual similarity of image keywords, and the visual similarity of image content. We assume that each query consists of a visual example and one or several keywords. Such queries naturally appear in web searching – a user may e.g. employ a standard keyword-based retrieval to search for images, then select a suitable visual representant and continue searching with both modalities.

4.3 Selection of Eligible Techniques

Having established the desired characteristics of a successful web image retrieval technique, we can use them to filter out unsuitable approaches and select promising techniques for further examination. Due to the flexibility requirement, we can directly rule out most early fusion solutions and focus on the late fusion techniques, which are by design suited for flexible searching. As discussed in Section 3.2, late fusion is realized during query evaluation, which may be composed of multiple phases. In this paper, we limit our attention to the two central ones that are crucial for the overall effectiveness and efficiency: basic search and postprocessing. Basic search is inevitably a part of each retrieval solution, while postprocessing is the most frequent strategy of result refinement. Considering these two phases and the two possible approaches to modality combination (see Section 3.1), we obtain four basic search strategies depicted in Figure 3: symmetric basic-search fusion realized by TA, asymmetric basic-search fusion that exploits some specialized index structure, symmetric postprocessing that approximates TA, and standard asymmetric postprocessing.

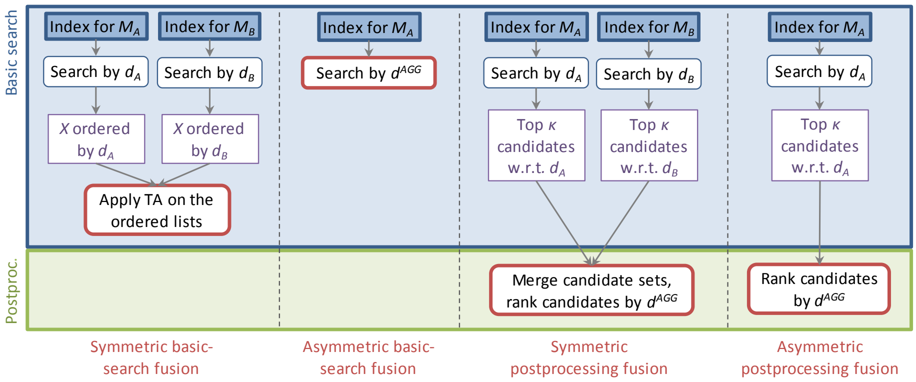
Basic search
Postproc.
Index for M_A
Index for M_B
Search by d_A
Search by d_B
X ordered by d_A
X ordered by d_B
Apply TA on the ordered lists
Index for M_A
Search by d^{AGG}
Index for M_A
Index for M_B
Search by d_A
Search by d_B
Top κ candidates w.r.t. d_A
Top κ candidates w.r.t. d_B
Merge candidate sets, rank candidates by d^{AGG}
Index for M_A
Search by d_A
Top κ candidates w.r.t. d_A
Rank candidates by d^{AGG}
Symmetric basic-search fusion
Asymmetric basic-search fusion
Symmetric postprocessing fusion
Asymmetric postprocessing fusion

Fig. 3. Possible late fusion scenarios.

In most of the existing image search systems, the choice between these options was based on implementation convenience and rather vague assumptions about the efficiency and effectiveness of these methods. Even though some of these systems provide good results [44, 89], the lack of rigorous performance evaluations makes it difficult to decide what factors participate in the success. Therefore, we decided to conduct a series of experiments that would allow us to quantify the performance of individual techniques and assess their usefulness.

5 Experimental Framework

To perform such a broad range of experiments and provide as fair comparison as possible, we have decided to implement all the retrieval methods using the same framework and run the experiments on the same hardware. In this section, we provide more information about the selected modalities that we have compared, the specific indexing techniques that have been utilized for efficient retrieval, and the parameters of the given techniques that have been examined.

5.1 Specification of Modalities

As stated earlier, a modality $\mathcal{M}$ is defined by a projection function $p_{\mathcal{M}}$ and a distance function $d_{\mathcal{M}}$. For the textual modality $\mathcal{M}^T$, we follow the traditional text retrieval paradigm – $p_{\mathcal{M}^T}$ extracts the keywords from a given multimodal object and performs standard normalization (stemming, stopword-filtering), $d_{\mathcal{M}^T}$ computes the cosine distance with *tf-idf* weighting [6]. The visual modality $\mathcal{M}^V$ can be represented by various types of global or local visual descriptors. Currently, global descriptors produced by deep convolutional neural networks (e.g. DeCAF [29]) represent state-of-the-art for visual similarity evaluation. However, at the time of our experiments the extraction of such descriptors for large data collections was prohibitively expensive. Therefore, we use the MPEG-7 [62] global descriptors in our experiments, which represent a reasonable compromise between retrieval quality and extraction costs. In particular, we employ a fixed combination of five MPEG-7 visual descriptors (Color Layout, Color Structure, Scalable Color, Edge Histogram, Homogeneous Texture) together with a distance function that is computed as a weighted sum of the partial distances evaluated by individual descriptors (for details, see [55]). The MPEG-7 descriptors are fused prior to data indexing (early fusion) and are regarded as a single visual descriptor in further discussions.

5.2 Implementation of the Fusion Methods

Most of the proposals of techniques for information retrieval are accompanied by experimental evaluations, therefore some implementation of the techniques can be acquired from their authors. However, the quality and reusability of the implemented prototypes vary greatly, the efficiency is heavily affected by the programming language that was employed, and the input and output data formats are usually specific to a given technique. On the other hand, in our complex experimental settings the most fair measure of efficiency is the wall-clock time. Therefore, we have decided to implement all the necessary techniques using our Java-based framework MESSIF [10], for which we already have several good implementations of state-of-the-art retrieval techniques.

For the visual similarity search, we have used the M-index [66] technique – a dynamic disk-based indexing approach that employs pivot-permutation approach along with various forms of metric dataspace pruning techniques to achieve online response times even on datasets with tens of million objects. For the text search, we have adopted the Lucene search engine [59] that was embedded in the MESSIF framework. Lucene provides fast text retrieval using classical tf-idf paradigm with several effectivity enhancements. Lucene is also able to provide online responses for tens of million indexed documents, which is our target dataset size.

In order to evaluate a multi-modal search, the aggregation function for combining the respective modalities must be provided. The modular design of the MESSIF library allows to plug in any user-specified function for the computation. In our case, we use a function that first normalizes the partial distances of each candidate object from a given modality and then uses a weighted sum to combine the partial distances. The influence of a given modality thus can be adjusted by providing weights for the summation. However, providing correct weights can be difficult for a user, therefore the weights can also be automatically learned on a sample collection for which a ground-truth is known.

Following from the analysis of large-scale image search task needs in Section 4, our comparison of multi-modal aggregation focuses on late fusion techniques. Specifically, we consider the symmetric basic-search fusion, asymmetric basic-search fusion, symmetric postprocessing fusion, and asymmetric postprocessing fusion. In case of asymmetric solutions, we consider that both $\mathcal{M}^T$ and $\mathcal{M}^V$ can be used as the primary modality. The following paragraphs detail the implementation of individual techniques:

Symmetric basic-search fusion This fusion technique is implemented by the standard Threshold Algorithm [33]. The indexes $I_{\mathcal{M}^T}$ and $I_{\mathcal{M}^V}$ provide the sorted access for the TA input while the MESSIF storage module provides a fast random access for retrieving the missing features need for computing the multi-modal distance d^{AGG}. The storage serves the data from a disk using a B-tree index built for the object identifiers.

Asymmetric basic-search fusion This approach is implemented by the *inherent fusion* technique [15], an approximate asymmetric late fusion method that computes the aggregated similarity of objects directly during the selection of the candidate set CS^{BS}. In particular, if the objects stored in a mono-modal index $I_{\mathcal{M}_1}$ contain feature descriptors for all other modalities (even though they are not used to build the index itself), the MESSIF library allows the user to alter the index searching procedure so that objects to be visited are identified by $d_{\mathcal{M}_1}$ but all visited objects are ranked directly by d^{AGG}. The number of objects to be visited is determined by the approximation parameter κ. Although no quality guarantees are given, this approach allows to compute the query distance function and the final candidate ranking efficiently for a large set of candidates identified by $d_{\mathcal{M}_1}$. In contrast to postprocessing asymmetric late fusion, the candidate identification and ranking can be run in parallel and thus much larger set of objects (typically by several orders of magnitude) can be visited within the same time limit.

Symmetric postprocessing fusion The postprocessing symmetric late fusion utilizes two sets of candidate objects, $CS^{BS}_{\mathcal{M}^V}$ and $CS^{BS}_{\mathcal{M}^T}$, retrieved from separate mono-modal indexes $I_{\mathcal{M}^T}$ and $I_{\mathcal{M}^V}$. Both $CS^{BS}_{\mathcal{M}^V}$ and $CS^{BS}_{\mathcal{M}^T}$ contain $\kappa/2$ objects, so that κ objects altogether are visited in the postprocessing phase. The objects from both basic-search results are merged and the d^{AGG} is computed using the MESSIF random-access storage. The top k ranking objects are then reported. This is in fact an approximation of the TA algorithm, where the sorted accesses are not incremental but provided completely as bulks. Note that the threshold constraint might not be satisfied, since the inspection of the sorted access lists might not be deep enough. Therefore, the effectiveness might be lower but the execution is much faster.

Asymmetric postprocessing fusion Finally, standard re-ranking is used to represent the asymmetric postprocessing fusion. The candidate set CS^{BS} of size κ is retrieved by one mono-modal index, and the candidates are ranked by the d^{AGG}. In comparison with the inherent fusion, this approach computes the multi-modal ranking only after the whole result is returned by the primary index.

6 Evaluation Plan

The objective of the experimental study is to evaluate both the efficiency and effectiveness of selected image retrieval methods in uniform conditions. In particular, we are interested in the following aspects: 1) applicability of precise retrieval techniques in large-scale searching, 2) effect of approximation on user satisfaction, 3) selection of method(s) with the best relevance-cost trade-off, and 4) identification of factors that influence the retrieval quality. With these objectives, we have designed the experiments as follows.

6.1 Datasets, Queries and Ground Truth

Even though the need for common benchmarking platforms is well recognized, there are only few datasets that can be used for image search evaluation [26]. In our particular case, we need a large collection of image-and-text data, accompanied with a ground truth for general multi-modal retrieval. To the best of our knowledge, no such testbed is publicly available apart from the Profiset platform[1] that we have introduced recently to enable large-scale retrieval evaluations [16]. The Profiset collection contains 20M high-quality images provided by the Profimedia photostock site[2]. Each image is accompanied by a rich and mostly error-free keyword annotation in English. Furthermore, the Profiset provides a set of 100 test queries, each of which is composed of a single example image and a short keyword description. The topics comprise a selection of the most popular queries from Profimedia search logs and several queries that are known to be either easy or difficult to process in content-based searching. A few examples are shown in Figure 4.

The Profiset platform does not provide a complete ground truth for the test queries, but offers tools for collecting a *partial ground truth*, i.e. relevance

[1] http://disa.fi.muni.cz/profiset

[2] http://www.profimedia.com

assessments for selected result objects. Our partial ground truth has been formed as follows: each result found by any tested method has been evaluated by at least two human judges, who have marked it as *highly relevant*, *partially relevant*, or *irrelevant* with respect to a given query. These categories have been transformed into relevance percentage (100 %, 50 %, and 0 %, respectively) and averaged, thus forming the relevance value of the given result object. More details about the ground truth collection process can be found in [16].

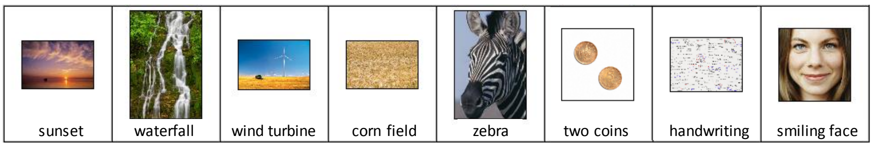

Fig. 4. Query objects.

Although the Profiset provides a suitable test environment, the evaluation results may be biased by the particular properties of this dataset. Therefore, we also employ a second testbed obtained from a different type of application. The CoPhIR test collection[3] contains images downloaded from the Flickr web gallery, accompanied by user-provided tags of unguaranteed quality. A 20M subset of the CoPhIR collection was randomly selected to make it comparable with the Profiset. The same set of test images and ground truth collection process have been used as with the Profiset testbed.

6.2 Performance and Quality Measures

For each experiment E, we measure the evaluation costs and the quality of the retrieved set of objects. Since all experiments are run in identical conditions, we can utilize wall-clock time as the measure of costs. Result quality is evaluated on both *distance-based* and *user-perceived* level (see Section 3.4). Let R_E be the result set returned in experiment E. The distance-based metric *relative error on distance at k (rED(k))* [99] compares R_E to a precise result R_{TA} provided by TA (as discussed in Section 3, it can be proved that TA finds the best k objects in terms of distance-based precision). Specifically, *rED(k)* compares the distances of objects at k-th position (d^k) in R_E and R_{TA}: $rED(k) = d^k_{R_E}/d^k_{R_{TA}} - 1$. The user-perceived quality is measured by the *Normalized discounted cumulative gain at k* ($NDCG(k)$), computed as a sum of user-provided relevance values of the k best objects from R_E normalized by their rank [42]. The $NDCG$ metric is applied in two modes: *natural* $NDCG$ ($NDCG_N$) is computed using multi-valued relevance assessments provided by users, whereas for *strict* $NDCG$ ($NDCG_S$) the relevance assessments are transformed into binary values so that only the results denoted as *highly relevant* are considered relevant. $NDCG_S$ thus represents a more demanding user.

[3] http://cophir.isti.cnr.it/

To guarantee maximum fairness, all experiments were run on the same single machine with 8 CPU cores and 32 GB RAM. In order to assess the scalability, some of the experiments were restricted to a single CPU.

6.3 Retrieval Parameters Settings

Method name	Fusion type	Approx. parameter κ	Abbreviation
Text search			Text
Content-based search			Visual
Threshold Algorithm	Symmetric, basis-search		TA
Approximate TA	Symmetric, postprocessing	100, 500, 2000, 30000	TA100, TA500, TA2K, TA30K
Visual-based inherent fusion	Asymmetric, basic-search	30000, 100000	VIF30K, VIF100K
Visual search with re-ranking	Asymmetric, postprocessing	100, 500, 2000	VR100, VR500, VR2K
Text search with re-ranking	Asymmetric, postprocessing	100, 500, 2000, 30000	TR100, TR500, TR2K, TR30K

Table 2. Overview of tested methods.

The parameters of the specific techniques, as described in Section 5.2, are summarized in Table 2. Mono-modal text search and content-based search constitute the baselines. For the approximate techniques, different values of the approximation parameter κ were tested. We should also notice that for asymmetric fusion with text as the primary modality, inherent fusion was not applied. The reason is that text-based retrieval is very efficient, thus re-ranking with $\kappa = 30000$ can be used to perform text-based fusion comparable (in terms of approximation strategy) to visual-based inherent fusion.

As discussed earlier, a crucial requirement for wide-domain searching is fusion flexibility, which is supported by all the above-listed methods. However, to avoid unnecessary confusion we only consider a single d^{AGG} in the experiments. A fixed weighted sum of the text- and visual-induced distances is used, the respective weights being chosen in a separate set of experiments so that average result quality is maximized.

7 Evaluation Results

As stated earlier, one of the main objectives of the experimental evaluation is to compare the precision and costs of various approaches to multi-modal image

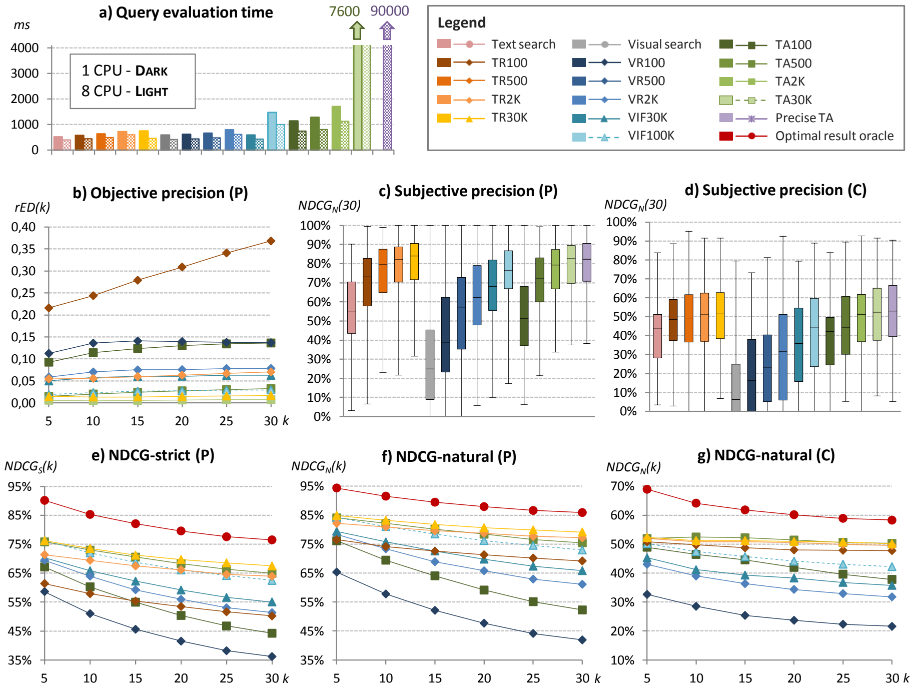

Fig. 5. Retrieval costs and precision for different search methods. (*Note that the graphs are color-coded.*)

retrieval, and to select the most suitable solution for large-scale image searching. Accordingly, the global effectiveness and efficiency of individual methods and the trade-off between these two characteristics are analyzed in the first part of this section. Afterwards, we present an additional analysis of the collected data that examines several factors specific for combining text and visual modalities of images.

7.1 Effectiveness & Efficiency: Overall Trends

An overall evaluation of experimental results in terms of both retrieval costs and result relevance is provided in Figure 5. In the following subsections, we examine individual graphs to answer the questions formulated in Section 6.

Applicability of precise retrieval The costs graph in Figure 5a clearly shows that precise late fusion, depicted in the rightmost bar, is not applicable in interactive large-scale searching. The response times are extremely high for TA, which is the only precise late fusion solution suitable for text and visual modalities that allows flexible query evaluation. The TA costs depend on the number of objects that need to be accessed, for which there is no theoretical bound. Our experiments have confirmed that for real data, such as ours, the number of visited objects is indeed very high.

Influence of approximation Since precise evaluation is too costly, an approximation is a vital concept to apply for interactive retrieval. Fortunately, we have been able to verify the generally accepted assumption that a certain level of distance-based imprecision does not result in any noticeable deterioration of result quality as perceived by users, which is illustrated by Figures 5b-g. However, it is important to select the approximation parameters appropriately.

Individual graphs in Figure 5 depict different aggregated metrics of result relevance measured over Profiset (P) and CoPhIR (C) collections. Figure 5b shows the objective result precision measured by rED, other graphs display the $NDCG$ metric of users' satisfaction. Quartile distribution of $NDCG$ for $kNN(30)$ query is shown in Figures 5c-d, which provide comparison of effectiveness for all tested methods. Figures 5e-g illustrate the development of average result relevance for various result sizes, using only a selection of methods to maintain readability. We can observe that there is an almost perfect agreement in the ordering of methods by rED and $NDCG$, with a single notable exception of text-based asymmetric fusion with a very rough approximation. This method provides poor results in terms of rED, but is considered rather good by users. This phenomenon is probably caused by users' preference for semantic relevance, which will be discussed more thoroughly later.

Considering the approximations, we can observe that for small candidate set size κ the quality of results is considerably worse than that of TA, especially when visual modality is used as the primary one. However, with sufficiently large κ the approximate techniques are comparable to or even slightly better than TA in terms of user satisfaction. The observed dependence between the result quality and candidate set size is roughly logarithmic. The largest improvements can be seen for visual-based asymmetric fusion where the trends suggest that even better results could be achieved if κ was higher. For text-based and symmetric fusion it seems that the optimum κ has been reached.

From the efficiency point of view, there are no large differences between the costs of asymmetric text-based and visual-based methods with the same κ, whereas the TA-based solutions are considerably more expensive. This is caused by the need to access two independent index structures which are not optimized for mutual cooperation. Subsequently, we do not consider symmetric fusion to be applicable for κ larger than a few thousand objects. On the other hand, both asymmetric variants are capable of processing 30000 candidate objects in about 0.5 s on moderately strong hardware, which we consider to be perfectly acceptable.

Optimal method selection The relevance results reveal that general trends of fusion effectiveness are very similar for both tested collections. All approximate late fusion techniques under consideration – text-based asymmetric fusion, visual-based asymmetric fusion, and approximate TA – are capable of achieving comparable result quality. However, the text-based asymmetric fusion slightly outperforms the other approaches in all quality measures, and requires less processing time to achieve a given level of relevance. This clearly makes it the most eligible method for both our datasets. Approximate TA comes as a close second in terms of result quality at any fixed approximation level, but its costs are prohibitive for the more precise variants. Visual-based approaches need to examine significantly more objects to achieve the same level of relevance, which is partly balanced by efficient inherent fusion implementation (Figure 5a).

The success of text-based methods is not surprising for the Profiset collection, which contains high-quality image annotations. We have also expected the text-based fusion to be less effective for the CoPhIR dataset than for Profimedia, as the quality of textual information in CoPhIR is significantly lower. However, we have assumed that the other approaches would be less influenced by the change of datasets. The absolute relevance values for text-based fusion are indeed about 30 % lower as compared to the Profiset, but the same applies for all results and the ordering of methods with respect to retrieval precision remains unchanged. Currently, we see at least three possible causes of such behavior: 1) human perception of relevance is more semantically-oriented than visually-oriented, therefore a text search result not visually similar to the query is more likely to be regarded relevant than vice versa (this corresponds to the observation made about the disproportion in rED and $NDCG$ relevance evaluations); 2) visual content descriptors and distance measures that were applied are not mature enough to capture the features important for users; and 3) the CoPhIR collection exhibits worse quality than the Profiset not only in the textual component but also in the quality of photos – in terms both of technical aspects (blur or other types of imaging noise) and relevance of content (many of the CoPhIR photos are difficult to interpret, do not attract the user, etc.). We believe that the observed results are influenced by all these factors, however a more detailed model than ours would be needed to determine their roles more precisely.

The fact that semantics is very important to users can also be observed in the difference between the evaluations by strict and natural $NDCG$. The ranking of methods by $NDCG_N$ is slightly different that ranking by $NDCG_S$. In particular, the highly approximate solution $TR100$ ranks higher by $NDCG_N$ than by $NDCG_S$, which supports our hypothesis that users appreciate semantic relevance even if the visual component is not sufficiently close to the query.

7.2 Uncovering Deeper Roots of Relevance

Even though the general findings in the previous section appoint the text-based asymmetric fusion as the most suitable search method, this approach is not optimal for all queries. In fact, about 40 % of queries would be better answered by a different method for Profiset and 50 % for CoPhIR. The red line denoted as "optimal result oracle" in Figures 5e-g shows the relevance level that could be achieved if the optimal method was chosen for each individual query. Unfortunately, it is very difficult to decide which fusion technique is the most suitable one for a given query. Our data shows that the effectiveness of methods differs from query to query, and often even for the same query when evaluated over two different datasets. To gain more insight into the behavior of multimedia retrieval, we study the experimental results from several less traditional perspectives.

Semantic categories The first aspect we examine are the query topics. In our experience, simple content-based retrieval works well on concepts like "sunset", "clouds", and other natural scenes, therefore we assume that some correlations could exist between the query topic and the effectiveness of individual fusion techniques (e.g. visual-based asymmetric fusion is expected to work well for pictures of nature, text-based for activities). Therefore, we have defined several categories that comprise popular search topics, and sorted our query objects into them. The less typical queries which do not fit into any category are not considered now.

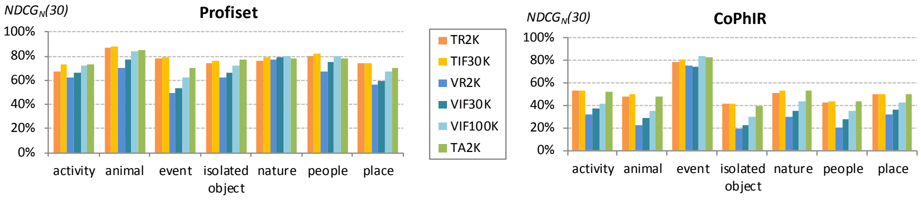

Fig. 6. Result relevance in different categories.

In Figure 6, we can see the effectiveness of selected methods for individual categories. The results do not show the expected relationships between categories and fusion behavior – although "nature" queries seem to be more visually-oriented than others in Profiset, this is not confirmed by CoPhIR. We conclude that semantic categories alone cannot be used to decide which fusion method to apply.

However, sorting the queries into categories has revealed other interesting details. Queries from the "event" category are better answered in CoPhIR than in Profiset, which contradicts the general observations presented in Section 7.1. We hypothesize that the observed phenomenon is caused by a higher occurrence of event-related images in CoPhIR, which was obtained from a photo-sharing site that is likely to contain such photos. The data indicate that the popularity of a given topic in the target database plays an important role for both the overall result quality and the applicability of individual search methods – whereas text-based search is clearly dominant for event queries in Profiset, visual-based fusion provides better results in CoPhIR.

Text-based relevance As shown in Section 7.1, methods that rely primarily on the text retrieval outperform the visual-based ones in a general case. To better understand the cases when the text modality misses relevant objects that the visual one is able to provide, we study the influence of the density of text descriptions on the retrieval effectiveness.

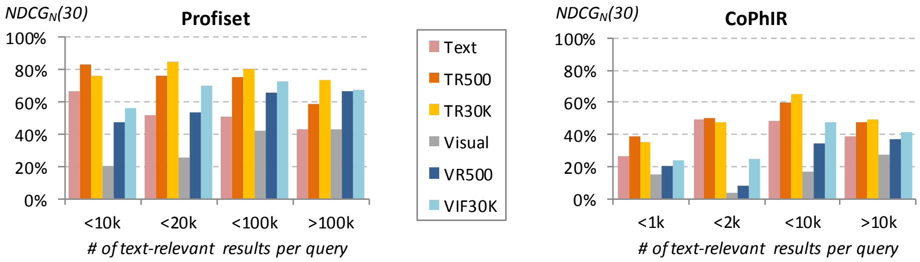

Fig. 7. Effectiveness versus text selectivity.

We have divided all query objects evenly into 4 groups based on the selectivity of the query text. Figure 7 shows the effectiveness of the selected methods in both the Profiset and CoPhIR collections, expressing the selectivity on x-axis by the total number of results possibly matching the query text. Note that the CoPhIR collection has only about a tenth of potential results as compared to the Profiset, which is caused by sparser annotations of the CoPhIR images. We can observe that for queries with more discriminative text (lower number of potential results) the text retrieval is distinctly more successful than methods that use primarily the visual search. However, for broader-term queries the visual similarity is becoming more important, e.g. for the group of queries with the lowest text selectivity – matching from 100,000 to 1.3 million objects – the visual-based asymmetric methods provide the same quality of results as text-based.

Method name	# of no-text results	# of relevant
Visual	136	27
VR500	98	12
VIF30K	25	3
TA	1	0
TA500	63	7
TA30K	4	0

Table 3. Retrieval of objects without text.

Next, we focus on the case where the text modality is not present in the target objects, thus the potentially relevant objects cannot be found by text search. This is only observable in the CoPhIR dataset where 28 % of objects contains only an automatic file name or no text at all. In the well-annotated Profiset collection, there are no images without text. Table 3 provides the numbers of results with no textual information returned by a respective method along with a number of

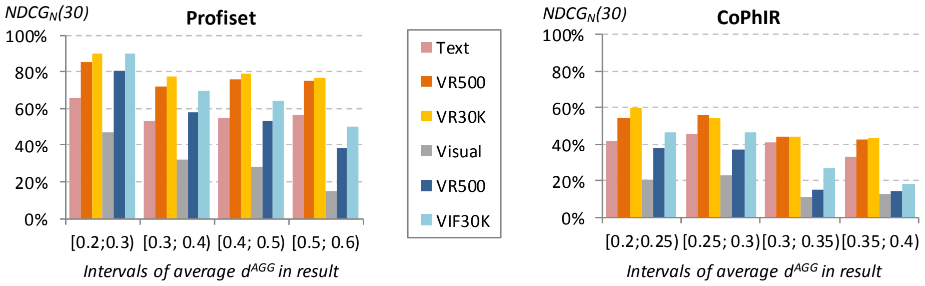

Fig. 8. Result relevance in relation to distance.

those results that were considered (highly) relevant by users. Note that methods based on text search are not included since they cannot find such objects. We can observe that most no-text objects are retrieved by the visual-only method followed by the visual search with text re-ranking. As expected, the number of results without text decreases for larger candidate sets, since the no-text objects are penalized by text-ranking. Interestingly, the visual-based methods are able to find relevant text-free results in about two thirds of the evaluated queries, and these results represents up to 6 % of the total returned relevant results.

Visual-based relevance Having analyzed the strengths and weaknesses of the text modality, let us now focus on the visual. Visual modality is inherently more problematic than text because of the semantic gap problem, and the fact that there is no strict differentiation between potentially relevant and irrelevant objects. However, these problems get reduced as the search space becomes denser. The improvement of visual search results with growing dataset size has been empirically observed in [8]. Google uses near-duplicate search for image annotation [91], which proves that in the web-scale searching the space is already dense enough. Since our previous analysis shows that text-based searching is reaching its limits for discriminating relevant objects in dense collections, we believe visual-based (or eventually also TA-based) solutions are better suited for such situations.

Although the density of our datasets is not high enough for the described phenomenon to apply, we can see some indications in Figure 8 that the distribution of objects in the search space is important. The average relevance of results obtained by approaches based on visual search reaches its maximum for queries that have low average distances of result sets, which corresponds to a higher density of the search space in the neighborhood of the respective query.

8 Summary and Discussion

The presented study is devoted to both theoretical and practical aspects of multimodal retrieval. In the first part, we have laid formal foundations for a systematic study of fusion techniques, and presented a new, comprehensive categorization of existing solutions. In the second part, we have focused on interactive large-scale image retrieval with text and visual modalities. In this context, we have compared two mono-modal search methods and four multi-modal late fusion techniques with different settings. The evaluation has been performed on two real-world datasets that are orders of magnitude larger than data usually employed in fusion evaluations. In particular, user-perceived relevance of more than 170,000 query-result pairs has been manually evaluated. This data allows us to study various aspects of image retrieval, including effectivenes, efficiency, and scalability. Let us now summarize the most important findings.

Our first conclusion concerns the applicability of individual late fusion solutions in large-scale multi-modal searching. We have found that precise flexible fusion is extremely costly on real-world data, while results of the same user-perceived quality can be obtained by efficient approximate solutions. Approximate solutions are thus more suitable. To maximize the chance of obtaining high-quality results, the approximation parameter κ should be chosen as high as efficiency limits allow. The observed dependence between the result quality and the candidate set size is roughly logarithmic.

For text-and-visual datasets of size and quality comparable to ours, text-based asymmetric fusion is very likely to provide optimal results in the majority of cases. Text-based searching is very strong, since it expresses semantics, it is also highly discriminative (there is a clear distinction between relevant and not-relevant objects), and the text searching is fast. Moreover, our experimental data shows that users tend to be satisfied with semantically relevant results even if the visual component is not sufficiently close to the query. If the quality of text information in a given collection is known to be low, symmetric late fusion stands as the most suitable solution, as it can best balance the strengths and weaknesses of both the modalities. However, the approximation then needs to be more rough because the processing costs are higher.

The data from our experimental evaluation also allowed us to study the suitability of individual fusion methods for different queries. While the text-based asymmetric fusion performed best on average on our data, it was far from being optimal for all queries. For nearly half of the queries, the result quality would be better if a different fusion method was chosen. A typical example are queries for which there are too few or too many text-relevant results, which would be better answered by visual-based asymmetric fusion. Our analysis discovered that the suitability of any given method is determined by both the specific query and the dataset properties. In particular, we studied the following aspects:

- Semantic categories: Classifying queries into semantic categories such as *nature*, *object*, etc. is alone not sufficient to decide which fusion method to apply. However, our data suggest that the popularity of a given topic within the dataset could be used to assess the suitability of fusion methods. In particular, the more popular topics tend to form dense subspaces that can be better searched by visual-based asymmetric fusion.
- Query text selectivity: As mentioned earlier, the text similarity is a highly effective tool for identifying candidate objects as long as the number of text-

relevant objects is not too high or too small. In case of broad-term queries with many relevant results the text prefiltering may not select the best candidate set. On the other extreme, objects without text descriptions cannot be found by text-based methods.
– Visual selectivity: We have discovered that visual-based fusion methods are most suitable for queries that have low average distances between objects in the result sets, which corresponds to a higher density of the search space in the neighborhood of the respective query.

These observations could be used in future to improve the quality of multimodal searching by dynamically choosing an optimal fusion method for a given query and dataset. The decision process would be based on statistics about the dataset and the properties of a given query. There is an intuitive parallel between such fusion optimization process and the query optimization performed by standard relational databases systems – in both cases, data statistics is used to estimate the best approach for query processing. However, the RDBS optimization aims at reducing the evaluation costs, whereas in fusion optimization we are mainly concerned with result quality. Based on the above-listed observations, we propose to utilize the following information as an input for the fusion optimization:

- Statistics of semantic categories: Using a suitable ontology of query topics and state-of-the-art classification methods, the dataset can be preprocessed so that individual objects are sorted into (possibly overlapping) semantic categories. On top of these, different statistics can be collected, including the category size, its visual density, and the number of no-text objects.
- Statistics of query text selectivity: Information about the selectivity of frequent text queries can be collected in advance or gradually at runtime.

To the best of our knowledge, the fusion optimization has not been considered before. It offers several new problems that can be studied in future, e.g. proposing algorithms for optimal fusion strategy selection, collecting and maintaining the statistics, or further data analysis to determine additional useful data properties.

Acknowledgments

This work was supported by the Czech national research project GA16-18889S. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085.

References

1. Abu-Shareha, A.A., Mandava, R., Khan, L., Ramachandram, D.: Multimodal concept fusion using semantic closeness for image concept disambiguation. Multimedia Tools and Applications 61(1), 69–86 (Jan 2011)

2. Ah-Pine, J., Csurka, G., Clinchant, S.: Unsupervised visual and textual information fusion in CBMIR using graph-based methods. ACM Transactions on Information Systems 33(2), 9:1–9:31 (2015)
3. Andrade, F.S.P., Almeida, J., Pedrini, H., da Silva Torres, R.: Fusion of local and global descriptors for content-based image and video retrieval. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP 2012). pp. 845–853 (2012)
4. Arampatzis, A., Zagoris, K., Chatzichristofis, S.A.: Dynamic two-stage image retrieval from large multimodal databases. In: 33th European Conference on IR Research (ECIR 2011). pp. 326–337 (2011)
5. Atrey, P.K., Hossain, M.A., El-Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems 16(6), 345–379 (2010)
6. Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval - the concepts and technology behind search, Second edition. Pearson Education Ltd., Harlow, England (2011)
7. Barrios, J.M., Bustos, B.: Automatic weight selection for multi-metric distances. In: Proceedings of the 4th International Conference on Similarity Search and Applications (SISAP 2011). pp. 61–68 (2011)
8. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubsky, J., Zezula, P.: Building a web-scale image similarity search system. Multimedia Tools and Applications 47, 599–629 (2010)
9. Batko, M., Kohoutkova, P., Zezula, P.: Combining metric features in large collections. In: 24th International Conference on Data Engineering Workshops (ICDE 2008). pp. 370–377 (2008)
10. Batko, M., Novak, D., Zezula, P.: MESSIF: Metric similarity search implementation framework. In: Digital Libraries: Research and Development, First International DELOS Conference, Revised Selected Papers. pp. 1–10 (2007)
11. Benavent, X., Garcia-Serrano, A., Granados, R., Benavent, J., de Ves, E.: Multimedia Information Retrieval Based on Late Semantic Fusion Approaches: Experiments on a Wikipedia Image Collection. IEEE Transactions on Multimedia 15(8), 2009–2021 (2013)
12. Blanken, H., de Vries, A., Blok, H., Feng, L.: Multimedia Retrieval. Data-Centric Systems and Applications, Springer (2007)
13. Bossé, É., Roy, J., Wark, S.: Concepts, models, and tools for information fusion. Artech House, Inc. (2007)
14. Bozzon, A., Fraternali, P.: Chapter 8: Multimedia and multimodal information retrieval. In: Search Computing, LNCS, vol. 5950, pp. 135–155. Springer Berlin / Heidelberg (2010)
15. Budikova, P., Batko, M., Novak, D., Zezula, P.: Inherent fusion: Towards scalable multi-modal similarity search. Journal of Database Management 27(4), 1–23 (2016)
16. Budikova, P., Batko, M., Zezula, P.: Evaluation platform for content-based image retrieval systems. In: International Conference on Theory and Practice of Digital Libraries (TPDL 2011). pp. 130–142 (2011)
17. Budikova, P., Batko, M., Zezula, P.: Similarity query postprocessing by ranking. In: 8th International Workshop on Adaptive Multimedia Retrieval – Revised Selected Papers. LNCS, vol. 6817, pp. 159–173 (2011)
18. Bustos, B., Kreft, S., Skopal, T.: Adapting metric indexes for searching in multimetric spaces. Multimedia Tools and Applications 58(3), 467–496 (2012)
19. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computing Surveys 44(1), 1 (2012)

20. Chatzichristofis, S.A., Zagoris, K., Boutalis, Y.S., Arampatzis, A.: A fuzzy rank-based late fusion method for image retrieval. In: 18th International Conference on Advances in Multimedia Modeling (MMM 2012). pp. 463–472 (2012)
21. Chen, L., Cong, G., Jensen, C.S., Wu, D.: Spatial keyword query processing: an experimental evaluation. In: The Proceedings of the VLDB Endowment (PVLDB). pp. 217–228 (2013)
22. Chen, Y., Yu, N., Luo, B., wen Chen, X.: iLike: integrating visual and textual features for vertical search. In: 18th International Conference on Multimedia (ACM Multimedia 2010). pp. 221–230 (2010)
23. Ciaccia, P., Patella, M.: Searching in metric spaces with user-defined and approximate distances. ACM Transactions on Database Systems 27(4), 398–437 (2002)
24. Clinchant, S., Ah-Pine, J., Csurka, G.: Semantic combination of textual and visual information in multimedia retrieval. In: Proceedings of the 1st International Conference on Multimedia Retrieval (ICMR 2011). p. 44 (2011)
25. Cong, G., Jensen, C.S., Wu, D.: Efficient retrieval of the top-k most relevant spatial web objects. The Proceedings of the VLDB Endowment (PVLDB) 2(1), 337–348 (2009)
26. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40, 5:1–5:60 (2008)
27. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: A large-scale hierarchical image database. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009). pp. 248–255 (2009)
28. Depeursinge, A., Müller, H.: Fusion techniques for combining textual and visual information retrieval. In: ImageCLEF, The Kluwer International Series on Information Retrieval, vol. 32, pp. 95–114. Springer Berlin Heidelberg (2010)
29. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31th International Conference on Machine Learning (ICML 2014). pp. 647–655 (2014)
30. Dong, Y., Gao, S., Tao, K., Liu, J., Wang, H.: Performance evaluation of early and late fusion methods for generic semantics indexing. Pattern Analysis and Applications 17(1), 37–50 (Apr 2013)
31. Eickhoff, C., Li, W., de Vries, A.P.: Exploiting user comments for audio-visual content indexing and retrieval. In: 35th European Conference on IR Research (ECIR 2013). pp. 38–49 (2013)
32. Escalante, H.J., y Gómez, M.M., Sucar, L.E.: Multimodal indexing based on semantic cohesion for image retrieval. Information Retrieval 15(1), 1–32 (2012)
33. Fagin, R.: Combining fuzzy information: an overview. SIGMOD Record 31, 109–118 (2002)
34. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press (1998)
35. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Transactions on Multimedia 13(2), 303–319 (2011)
36. Ha, H., Yang, Y., Fleites, F., Chen, S.: Correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval. In: Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME 2013). pp. 1–6 (2013)
37. Hemayati, R.T., Meng, W., Yu, C.T.: Semantic-based grouping of search engine results using wordnet. In: Advances in Data and Web Management. pp. 678–686 (2007)

38. Hoque, E., Strong, G., Hoeber, O., Gong, M.: Conceptual query expansion and visual search results exploration for web image retrieval. In: 7th Atlantic Web Intelligence Conference (AWIC 2011). pp. 73–82 (2011)
39. Hörster, E., Slaney, M., Ranzato, M., Weinberger, K.: Unsupervised image ranking. In: 1st ACM workshop on Large-scale multimedia retrieval and mining (LS-MMRM '09). pp. 81–88 (2009)
40. Hsu, W.H., Kennedy, L.S., Chang, S.F.: Reranking methods for visual search. IEEE MultiMedia 14(3), 14–22 (2007)
41. Jain, R., Sinha, P.: Content without context is meaningless. In: International conference on Multimedia (ACM Multimedia 2010). pp. 1259–1268. ACM (2010)
42. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)
43. Jegou, H., Schmid, C., Harzallah, H., Verbeek, J.J.: Accurate image search using the contextual dissimilarity measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 2–11 (2010)
44. Jing, Y., Baluja, S.: VisualRank: Applying PageRank to large-scale image search. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(11), 1877–1890 (2008)
45. Khasanova, R., Dong, X., Frossard, P.: Multi-modal image retrieval with random walk on multi-layer graphs. In: IEEE International Symposium on Multimedia (ISM 2016). pp. 1–6 (2016)
46. Kherfi, M.L., Ziou, D., Bernardi, A.: Image retrieval from the World Wide Web: Issues, techniques, and systems. ACM Computing Surveys 36, 35–67 (2004)
47. Kludas, J., Bruno, E., Marchand-Maillet, S.: Information fusion in multimedia information retrieval. In: Adaptive Multimedial Retrieval: Retrieval, User, and Semantics, LNCS, vol. 4918, pp. 147–159. Springer Berlin / Heidelberg (2008)
48. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: 26th Annual Conference on Neural Information Processing Systems (NIPS 2012). pp. 1106–1114 (2012)
49. Lai, K., Liu, D., Chang, S., Chen, M.: Learning sample specific weights for late fusion. IEEE Transactions on Image Processing 24(9), 2772–2783 (2015)
50. Lan, Z.z., Bao, L., Yu, S., Liu, W., Hauptmann, A.: Double fusion for multimedia event detection. 18th International Conference on Advances in Multimedia Modeling (MMM 2012) pp. 173–185 (2012)
51. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. TOMCCAP 2(1), 1–19 (2006)
52. Li, J.: Reachability based ranking in interactive image retrieval. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015). pp. 867–870 (2015)
53. Li, J., Ma, Q., Asano, Y., Yoshikawa, M.: Re-ranking by multi-modal relevance feedback for content-based social image retrieval. In: 14th Asia-Pacific Web Conference on Web Technologies and Applications (APWeb 2012). pp. 399–410 (2012)
54. Liu, Y., Mei, T., Hua, X.S.: CrowdReranking: exploring multiple search engines for visual search reranking. In: 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009). pp. 500–507 (2009)
55. Lokoc, J., Novák, D., Batko, M., Skopal, T.: Visual image search: Feature signatures or/and global descriptors. In: 5th International Conference on Similarity Search and Applications (SISAP 2012). pp. 177–191 (2012)
56. Ma, D., Yu, Z.: New video target tracking algorithm based on KNN. Journal of Multimedia 9(5), 709–714 (2014)

57. Magalhães, J.a., Rüger, S.: An information-theoretic framework for semantic-multimedia retrieval. ACM Transactions on Information Systems 28(4), 1–32 (Nov 2010)
58. May, W., Fidler, S., Fazly, A.: Unsupervised disambiguation of image captions. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics (SemEval 2012). pp. 85–89 (Jun 2012)
59. McCandless, M., Hatcher, E., Gospodnetić, O.: Lucene in Action: Covers Apache Lucene V. 3. 0. Manning Pubs Co Series, Manning (2010)
60. Mei, T., Rui, Y., Li, S., Tian, Q.: Multimedia search reranking. ACM Computing Surveys 46(3), 1–38 (Feb 2014)
61. Mironica, I., Ionescu, B., Vertan, C.: Hierarchical clustering relevance feedback for content-based image retrieval. In: 10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012). pp. 1–6 (2012)
62. MPEG-7: Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002 (2002)
63. Müller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: Experimental Evaluation in Visual Information Retrieval. Springer, 1st edn. (2010)
64. Nga, D.H., Yanai, K.: VisualTextualRank: An extension of VisualRank to large-scale video shot extraction exploiting tag co-occurrence. IEICE Transactions on Information & Systems 98-D(1), 166–172 (2015)
65. Novák, D.: Multi-modal similarity retrieval with distributed key-value store. Mobile Networks and Applications 20(4), 521–532 (2015)
66. Novak, D., Batko, M., Zezula, P.: Metric index: An efficient and scalable solution for precise and approximate similarity search. Information Systems 36(4), 721–733 (2011)
67. Oh, S., McCloskey, S., Kim, I., Vahdat, A., Cannons, K.J., Hajimirsadeghi, H., Mori, G., Perera, A.G.A., Pandey, M., Corso, J.J.: Multimedia event detection with multimodal feature fusion and temporal concept localization. Machine Vision and Applications 25(1), 49–69 (2013)
68. Park, G., Baek, Y., Lee, H.K.: Web image retrieval using majority-based ranking approach. Multimedia Tools and Applications 31(2), 195–219 (2006)
69. Patella, M., Ciaccia, P.: Approximate similarity search: A multi-faceted problem. Journal of Discrete Algorithms 7(1), 36–48 (2009)
70. Pedronette, D.C.G., da Silva Torres, R.: Combining re-ranking and rank aggregation methods for image retrieval. Multimedia Tools and Applications 75(15), 9121–9144 (2016)
71. Pham, T.T., Maillot, N., Lim, J.H., Chevallet, J.P.: Latent semantic fusion model for image retrieval and annotation. In: Sixteenth ACM Conference on Information and Knowledge Management (CIKM 2007). pp. 439–444 (2007)
72. Pulla, C., Jawahar, C.V.: Multi modal semantic indexing for image retrieval. In: 9th ACM International Conference on Image and Video Retrieval (CIVR 2010). pp. 342–349 (2010)
73. Qi, S., Wang, F., Wang, X., Guan, Y., Wei, J., Guan, J.: Multiple level visual semantic fusion method for image re-ranking. Multimedia Systems 23(1), 155–167 (2017)
74. Richter, F., Romberg, S., Hörster, E., Lienhart, R.: Multimodal ranking for image search on community databases. In: Proceedings of the international conference on Multimedia information retrieval (MIR '10). pp. 63–72 (2010)
75. Rokach, L.: Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. Computational Statistics & Data Analysis 53(12), 4046–4072 (2009)

76. Ross, A., Jain, A.K.: Multimodal biometrics: An overview. In: 12th European Signal Processing Conference. pp. 1221–1224 (2004)
77. Rui, Y., Huang, T., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool for interactive content-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology 8(5), 644–655 (1998)
78. Safadi, B., Sahuguet, M., Huet, B.: When textual and visual information join forces for multimedia retrieval. In: International Conference on Multimedia Retrieval (ICMR 2014). p. 265 (2014)
79. Samet, H.: Foundations of Multidimensional and Metric Data Structures. Computer Graphics and Geometric Modeling, Morgan Kaufmann Publishers Inc. (2005)
80. dos Santos, J.M., Cavalcanti, J.M.B., Saraiva, P.C., de Moura, E.S.: Multimodal re-ranking of product image search results. In: Advances in Information Retrieval - Proceedings of the 35th European Conference on IR Research (ECIR 2013). pp. 62–73 (2013)
81. Santos Jr., E., Gu, Q.: Automatic content based image retrieval using semantic analysis. Journal of Intelligent Information Systems 43(2), 247–269 (2014)
82. Siddiquie, B., White, B., Sharma, A., Davis, L.S.: Multi-modal image retrieval for complex queries using small codes. In: International Conference on Multimedia Retrieval (ICMR 2014). p. 321 (2014)
83. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349 –1380 (2000)
84. Snoek, C., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: 13th ACM International Conference on Multimedia (ACM Multimedia). pp. 399–402 (2005)
85. Sugiyama, Y., Kato, M.P., Ohshima, H., Tanaka, K.: Relative relevance feedback in image retrieval. In: International Conference on Multimedia and Expo (ICME 2012). pp. 272–277 (2012)
86. Tollari, S., Detyniecki, M., Marsala, C., Fakeri-Tabrizi, A., Amini, M.R., Gallinari, P.: Exploiting visual concepts to improve text-based image retrieval. In: 31th European Conference on IR Research (ECIR 2009). pp. 701–705 (2009)
87. Tran, T., Phung, D., Venkatesh, S.: Learning sparse latent representation and distance metric for image retrieval. In: IEEE International Conference on Multimedia and Expo (ICME 2013). pp. 1–6. IEEE (2013)
88. Uluwitige, D.C.N.W., Chappell, T., Geva, S., Chandran, V.: Improving retrieval quality using pseudo relevance feedback in content-based image retrieval. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2016). pp. 873–876 (2016)
89. Wang, L., Yang, L., Tian, X.: Query aware visual similarity propagation for image search reranking. In: ACM Multimedia 2009. pp. 725–728 (2009)
90. Wang, W., Yang, X., Ooi, B.C., Zhang, D., Zhuang, Y.: Effective deep learning-based multi-modal retrieval. The VLDB Journal 25(1), 79–101 (2016)
91. Wang, X.J., Zhang, L., Ma, W.Y.: Duplicate-search-based image annotation using web-scale data. Proceedings of the IEEE 100(9), 2705–2721 (2012)
92. Wei, Y., Song, Y., Zhen, Y., Liu, B., Yang, Q.: Heterogeneous translated hashing: A scalable solution towards multi-modal similarity search. ACM Transactions on Knowledge Discovery from Data 10(4), 36:1–36:28 (2016)
93. Wilkins, P., Smeaton, A.F., Ferguson, P.: Properties of optimally weighted data fusion in CBMIR. In: 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). pp. 643–650 (2010)

94. Wu, P., Hoi, S.C.H., Zhao, P., Miao, C., Liu, Z.: Online multi-modal distance metric learning with application to image retrieval. IEEE Transactions on Knowledge and Data Engineering 28(2), 454–467 (2016)
95. Xiao, Z., Qi, X.: Complementary relevance feedback-based content-based image retrieval. Multimedia Tools and Applications 73(3), 2157–2177 (2014)
96. Xu, S., Li, H., Chang, X., Yu, S., Du, X., Li, X., Jiang, L., Mao, Z., Lan, Z., Burger, S., Hauptmann, A.G.: Incremental multimodal query construction for video search. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR 2015). pp. 675–678 (2015)
97. Yang, X., Zhang, Y., Yao, T., Ngo, C., Mei, T.: Click-boosting multi-modality graph-based reranking for image search. Multimedia Systems 21(2), 217–227 (2015)
98. Zezula, P.: Future trends in similarity searching. In: 5th International Conference on Similarity Search and Applications (SISAP 2012). pp. 8–24 (2012)
99. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search – The Metric Space Approach, Advances in Database Systems, vol. 32. Springer (2006)
100. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. Pattern Recognition 45(1), 346–362 (2012)
101. Zhang, S., Yang, M., Cour, T., Yu, K., Metaxas, D.N.: Query specific fusion for image retrieval. In: 12th European Conference on Computer Vision (ECCV 2012). pp. 660–673 (2012)
102. Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). pp. 1741–1750 (2015)
103. Zitouni, H., Sevil, S.G., Ozkan, D., Duygulu, P.: Re-ranking of web image search results using a graph algorithm. In: 19th International Conference on Pattern Recognition (ICPR 2008). pp. 1–4 (2008)

Response to the reviewers' comments

REVIEW #1 AUTHOR COMMENTS

A. General evaluation (Yes or No, with brief comments if necessary) *1. Does the title properly describe the paper? Yes 2. Does the abstract bring out the main points of the paper? Yes 3. Does the scientific content of the paper justify the space it will occupy? Yes 4. Is it of interest to the readership? Yes 5. Does the paper contain errors of fact or logic? No*

B. Detailed comments for the authors *The paper presents a comparative analysis of multi-modal data retrieval techniques, with a focus on their applicability for interactive large scale search. In the multi modal retrieval, multiple complementary views of the data objects (e.g, textual description and content of images) are combined to make the data retrieval more efficient and precise. The authors firstly describe a formal model of multi-modal retrieval. Then, they describe the two main categories of the existing techniques, namely early fusion and late fusion. Afterworlds, they analyze the applicability of the existing techniques for large scale image retrieval. The analytical study is followed by an extensive experimental evaluation of the multi modal technique on two real-world datasets. Overall, the paper is well written and the different techniques of multi modal data retrieval are well analyzed. The experimental evaluation is convincing, and the results are useful for the people from research or industry who are interested in a comparison between different multi-modal techniques for data retrieval over image datasets. For example, the results show that high quality results can be obtained by using efficient approximate solutions, thus no need to the exact techniques that are highly costly.*

Some comments:

- *In Section 3.2, the authors describe briefly the threshold algorithm (TA) as a symmetric basic-search-phase fusion. It would be worth to give more details about TA, for example its stopping condition, the sorted and random accesses used by this algorithm, etc.*

  We have extended the description of TA to include more details about the algorithm.

- *In addition, I would like to see how the different access methods of TA (particularly random access) have been implemented in the evaluation tests.*

  The implementation of TA is described in Section 5.2. In particular, the random access is implemented by the MESSIF [10] storage module, which serves the data from a disk using a B-tree index built for the object identifiers.

- *In Section 6.2, what do you mean by approximate TA, which is one of the tested models in the experiments? This method should be clearly defined.*

  The overview of methods in Section 6.2 refers to Section 5.2, where individual methods are described in detail. As "approximate TA" we denote symmetric postprocessing fusion.

– *Typo: Page 28: Queries from the ,,event” → Queries from the “event”.*

Thank you for noticing, the quotation marks have been corrected.

REVIEW #2 AUTHOR COMMENTS

This paper contains a systematically survey of the existing multi-modal search techniques. The authors analyze their strengths and describe some methods considered as representative. Then the paper describes an experimental evaluation of the techniques used in a large-scale multi-modal image retrieval on the web.

The paper seems to be written about two-three years ago. It is better for the authors to review their work before submitting for publishing. e.g. the URL http://mufin.fi.muni.cz/profiset provided in the paper does not work.

The paper was written more than two years ago. Most of the time, it has been stuck in the TLDKS processing queue – apparently, there was some problem with the upgrade of the submission system, during which our paper was forgotten. We have of course checked and updated both the links and the related work.

The Bibliography item 18. Budikova, P., Batko, M., Novak, D., Zezula, P.: Large-scale multi-modal image search: theory and practice. International Journal of Multimedia Data Engineering and Management (IJMDEM) (2014), accepted for publication. mentioned as accepted for publication in 2014, is not currently published or at least I could not find it.

The cited paper was indeed accepted for publication in IJMDEM in 2014. Unfortunately, after the acceptance letter something went wrong with IJMDEM – the paper was never published and the editors did not communicate with us. After some time, we gave up the efforts, rewrote the paper and published it in a different journal. This bibliography item was therefore replaced by the following: 15. Budikova, P., Batko, M., Novak, D., Zezula, P.: Inherent fusion: Towards scalable multi-modal similarity search. Journal of Database Management 27(4), 123 (2016)

I suggest to the authors to remove the URL links contained in the items from Bibliography - in this way the space occupied by that section will shrink.

Thank you for the suggestion, we have removed the URLs.

pag. 20 Symmetric basic-search fusion - ”The indexes I_M^T and I_M^T provide ...” I believe there is a mistake here!

Thank you for noticing, this was indeed a mistake, we have corrected it.

REVIEW #3 AUTHOR COMMENTS

This paper studies the problem of multi-modal image retrieval. It first provides a survey of this area, and then describes an experimental evaluation of different techniques. The literature review has to be updated with more recent works. The *more recent paper in the bibliography is 3 years old, while most of the referenced papers are 5-9 years old.*

The paper was written more than two years ago. Most of the time, it has been stuck in the TLDKS processing queue – apparently, there was some problem with the upgrade of the submission system, during which our paper was forgotten. Therefore, the related work was not up-to-date. We have updated it with a number of recent research works, while some of the older ones have been removed.

In particular, the authors should include, and experiment with, recent content-based image retrieval methods.

Recent advancements in the field of deep convolutional neural networks resulted in new, effective visual descriptors such as DeCAF. It would be interesting to utilize these in our experiments. However, at the time of our experiments, the extraction of these descriptors was extremely costly for large datasets. Therefore, we decided to utilize the MPEG-7 descriptors. The reasons of this choice are now mentioned in Section 5.1. Based on our more recent experience with the DeCAF features, we do not expect that additional experiments with these features would bring any significant new findings, while their evaluation would require a lot of time and effort, including manual labor necessary for quality assessments of any new results.

The taxonomy used to organize the related work is really useful, but should be improved by taking into consideration the following comments. The difference between symmetrical and asymmetrical fusion is not immediately clear, especially since symmetrical fusion also allows for the weighting of the various modalities. The authors should stress more the aspect that makes these two categories different (if I understand well, the way that the query answering algorithm works).

We have reformulated the second paragraph in Section 3.1, which now better explains the difference between symmetric and asymmetric fusion.

Sections 3.1 to 3.6 propose the organization of the various approaches along different axis, but do not explain if/how these axis can be combined. The authors should provide a detailed discussion on which of these axis are orthogonal, how they can be combined, and what the relevant considerations would be in each one of these cases.

It is mentioned in the beginning of Section 3 that the individual axis are not orthogonal and influence each other. The mutual influencing is then explained in the discussions of individual axes. To provide a complex view on the axis correlations, we have added Table 1 that summarizes the meaningful combinations of individual approaches.

The discussion of the conclusions is rather poor. The lessons learned that are included in the conclusions are the expected ones. The paper should spend more effort to discuss more surprising aspects and results of the evaluation, as well as the corresponding insights and intuitions.

We have significantly extended the discussion of conclusions. It now summarizes the most interesting experimental findings and outlines the possible use of these findings for the optimization of multi-modal search quality.

Figure 1 is so small that it is unreadable. All the font sizes should be significantly increased in size.

We have enlarged the whole Figure 1, putting it on a separate page.