Advanced Retrieval Models
for Web Image Search

Linjun Yang

Advanced Retrieval Models for Web Image Search

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universtiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College van Promoties,
in het openbaar te verdedigen op dinsdag 9 juli 2013 om 12:30 uur
door

Linjun YANG

Master of Science in Computer Science,
Fudan University, China
geboren te Hanshan, Anhui, China.

Dit proefschrift is goedgekeurd door de promotoren:

Prof.dr. A. Hanjalic
Prof.dr.ir. R.L. Lagendijk

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr. A. Hanjalic,	Technische Universiteit Delft, promotor
Prof.dr.ir. R.L. Lagendijk,	Technische Universiteit Delft, promotor
Prof.dr. A.P. de Vries,	Technische Universiteit Delft
Prof.dr. C. Witteveen,	Technische Universiteit Delft
Prof.dr.ir. W. Kraaij,	Radboud Universiteit Nijmegen
Dr. M.S. Lew,	Universiteit Leiden
Dr. S. Li,	Microsoft Research Asia

Microsoft Corporation has provided substantial support in the preparation of this thesis.

ISBN 978-94-6186-172-6

Contents

Chapter 1

Introduction

1.1 Introduction to Web image search

To facilitate access to the rapidly growing collections of images on the Web and maximize their benefit for the users, *image search* has become an increasingly important research topic. We can distinguish between two main schemes for searching for images on the Web. In the first, *keyword-based* scheme illustrated in Figure 1.1, images are searched for by a *query* in the form of a textual keyword. This scheme can be seen as a direct extension of the widely adopted general Web search. The second, *example-based* scheme allows the users to search for similar images by providing an uploaded example image serving as query. This scheme is illustrated in Figure 1.2. While it has been deployed more and more by commercial Web search engines (e.g. TinEye [7]), this scheme is also highly valuable for potential usage in a mobile search scenario (e.g. Google Goggles [5]).

Independent of which search scheme is deployed, an image search engine generally operates in two main steps: the offline *index generation* and the online *index serving* step. The main purpose of the index generation step, frequently referred to as *indexing*, is to improve the efficiency of image search and keep this efficiency scalable with the increasing size of image collections. In the indexing step, the images discovered and crawled from the Web are processed to generate the *metadata* (the "data about the data") that represent the content of the images in an informative and discriminative fashion. The metadata may include visual signal representation (*visual features*) of the images acquired using the image analysis techniques combing image processing and computer vision, but also manually inserted and automatically inferred textual annotations. Regarding the extraction of visual features, one of the most notable achievements was the development of the SIFT (Scale-Invariant Feature Transform) features [61] and the invention of image representation in the form of a *bag of visual words* (BoW) [93]

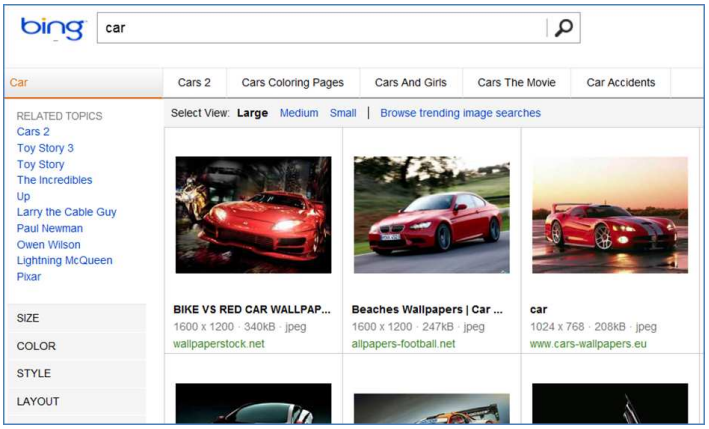

Figure 1.1: *Illustration of the keyword-based image search scheme. In this scheme, the search query comprises one or multiple keywords specified by the user.*

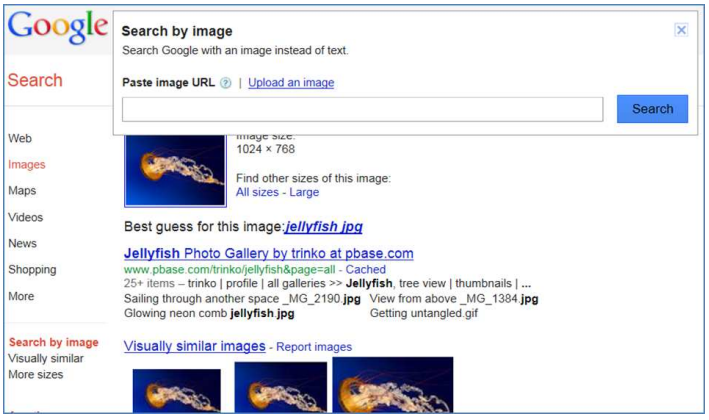

Figure 1.2: *Illustration of the example-based image search scheme. In this scheme, the query is an image, either specified by a URL or uploaded by users.*

that builds on SIFT features. A characteristic example of the manually inserted textual annotations are user-generated keywords, or *tags* that typically accompany the images users upload on the Web. Automatically inferred annotations are usually acquired by analyzing the (e.g. surrounding) texts on the Web pages hosting the images and deriving the most informative and discriminative keywords using the theory and algorithms of *information retrieval* [67]. Since metadata is extracted and formatted to be much more compact than the original image data, searching for images based on metadata is a key to an efficient interaction with a large image collection.

In the index serving step, the query submitted to the search engine is first transformed into a representation that is compatible with the metadata of the

images in the collection. While for keyword queries query alteration and expansion may be performed, image queries are processed using the same image analysis techniques as in the indexing step to extract their visual features. Then, the query representation serves as input into the *ranking* module where it is compared with the metadata of the images in the collection. The level of match provides the basis for estimating the *relevance* of an image to the query. The relevance is then used to rank the images in the collection. Finally, the ranked list of images serves as the response (*results list*) of the image search engine to the search request of the user.

1.2 Retrieval models

The ranking component of a general search engine deploys a *retrieval model* that suggests how to compute the relevance of a document to a query. Retrieval models are therefore critical for the success of a search engine.

In the keyword-based image search scenario, the basic retrieval models have been adopted from the general text-document search. Already for decades, such models have been among the most important topics of fundamental research in the information retrieval community [67]. The models including *tf-idf* [67], *Okapi BM25* [80], *language models* [137] and *learning-to-rank* methods [59] are among the most prominent retrieval models that have significantly influenced the development of search engines over the past years.

In the example-based image search scenario, also retrieval models are required to estimate the relevance of an image to the query image based on the match between the visual metadata and the features extracted from the query image. High efficiency and scalability of the BoW-based image representation has made it widely adopted as the basis for building a retrieval model. The modeling step itself, however, has typically been approached by extrapolating the models from the text domain mentioned above onto the visual domain. Here, in particular, the language modeling approach has been the most effective one [31].

1.3 Problem statement

While the aforementioned conceptual solutions for developing retrieval models have been widely adopted, their success in enabling an image search engine to provide a high-quality results list critically depends on how solid the foundations are, based on which metadata are extracted and compared with a query. We now briefly elaborate on the parameters influencing this success for both the keyword-based and example-based image search.

Wide adoption of the keyword-based Web image search scheme [1][2][3] is based on the rationale that Web images are usually hosted on Web pages. There, various texts, including the page title, professional or user-generated annotations and the

text surrounding the image on the page, can be considered generally available and potentially useful to index the images contained on the page. Following this rationale, images on the Web can be searched indirectly via the accompanying textual information that is matched with a textual search query, for the purpose of which proven techniques adopted from the general Web search can be used.

A typical problem encountered during the keyword-based image search is that the relevance link between the images found on a Web page and the surrounding text is not always obvious. The text (e.g. about politics) may be rich and its keywords may point to different possible categories of visual content (e.g. different politicians, interviews, journalists, people on the street discussing politics). Related to this, the limited number of images displayed on a Web page may only reflect some of its textual content. Consequently, not all metadata derived from the text and attached to the images would have the same relevance in respect to the visual content of the images. This leads to another, though related problem, namely that a textual query will typically lead to a large variety of visual content, the relevance of which may vary across search scenarios.

To address the mismatch problems illustrated above, significant research effort has been invested over the past years in developing solutions that automatically annotate images by the keywords derived directly from the visual content of the images. The keywords in this case are expected to be related to *visual semantic concepts* [96] that correspond to the objects, persons and scenes depicted in the images (e.g. "tree", "George W. Bush", "car", "landscape"). Image indexing based on semantic concepts [73, 96] essentially consists of two steps. First, a model is learned per semantic concept in a supervised fashion, and then, based on the model fit, the probability is estimated whether a given image contains a particular semantic concept. While this paradigm is theoretically effective in bringing the image content and textual keywords closer to each other in terms of relevance, the results reported within the TRECVID evaluation benchmark [18, 97] have shown that in practice only a limited success could be expected using this paradigm. The first problem lies in the insufficient capability of the paradigm to scale up to a large number of concepts that are required to cover a realistic query space [34]. The second problem lies in the semantic gap [96] between the generally high abstractness of the semantic concepts and the visual features used to train the concept models. This gap becomes even larger in case of more abstract categories of semantic concepts, like those *thematic* ones (e.g. "politics") that do not directly relate to the visual content of an image, but rather address the general thematic context the visual content of the image belongs to.

For the example-based image search scheme, visual metadata have the advantage to be directly related to the visual content of the images. Furthermore, the widely adopted solutions for image analysis (e.g. using SIFT) are robust and scalable. However, they too suffer from their own specific deficiencies. The main problem lies in the fact that the visual metadata typically do not reflect the relevance criteria users impose on the image search engine. While visual metadata may point to an image with e.g. particular bit-pattern distribution, users are typ-

ically interested in semantics - the meaning represented by these bit-patterns, for instance at the level of semantic concepts, either visual or thematic ones discussed above. Due to the underlying principle of matching visual features across images, image search using visual metadata will typically lead to insufficient diversity of the visual content in the results list.

In summary, we can state that metadata derived from the surrounding text may be semantically too complex to provide clear relevance links to images on the Web. Making the relevance links between the keywords and visual content stronger (e.g. by linking keywords to semantic concepts learned from the images) is, however, not scalable and not always feasible. On the other hand, visual metadata can be extracted in a scalable fashion, but they are in general insufficiently informative of image semantics to enable effective retrieval. It can also be stated that textual metadata make the relevance space too broad, while the visual metadata limit this space too much.

1.4 Thesis contribution

Building on the state-of-the-art in metadata extraction and relevance estimation and in view of the problems discussed above, this thesis proposes a number of novel insights and approaches for improving the retrieval models for both keyword- and example-based image search. While we adopt the standard solutions for text and image analysis, we investigate

- the possibilities to enrich the information used to estimate the image relevance to the query, and

- the methods to deploy this information to verify and enhance the search results obtained from metadata matching,

which should lead to *better informed* retrieval models.

We address the improvement of retrieval models in two ways, each of them covering one of the image search scenarios. In Part I of the thesis, we focus on the keyword-based image search and investigate how multiple information resources can be deployed to refine the initial results list through *reranking*. Then, in Part II, we explore the possibilities to exploit the *contextual* information to enrich the relevance model. Here both the contexts of query formulation and the image collection are considered.

In the remainder of this chapter, we elaborate in more detail on the rationale, scope and contribution of the material presented in each part of the thesis and explain the organization of the thesis material across the chapters.

1.5 Part I: Visual reranking for keyword-based image search

Image search reranking stands for the category of techniques that are devised to reorder (refine) the image search results list returned by the text search engine. The refinement aims at a new results list that has better overall relevance to the query than the original one. Since, typically, the information extracted from the visual content of the initially returned images is deployed to derive the reranking criteria, image search reranking is also often referred to as *visual (image search) reranking*. The essence of reranking is to find the optimal trade-off between the initial results list and the influence of the reranking criteria. In this way, the search benefits from the synergy of information derived from two *modalities* – visual and text, which is expected to make it more powerful than the pure text-based search.

Visual reranking has initially been introduced as an unsupervised paradigm since no supervised offline model learning was required to generate the reranking function. Instead, the reranking function is learned online based on predefined visual reranking criteria, like for example the requirement that visually similar images are positioned close to each other in the new results list [102]. Although the unsupervised nature of this reranking paradigm preserves the search scalability, this paradigm also suffers from problems that make it insufficiently effective for broad deployment in the image search practice. The existing reranking criteria are namely based on heuristic assumptions regarding the role of the visual modality in determining the relevance of the search results it is supposed to refine. Since these assumptions may not be valid to the same degree in different search scenarios, the reranking performance remains largely unpredictable.

In the first part of the thesis we provide insights and novel technical contributions for which we believe to help the research on visual reranking to effectively address the deficiencies mentioned above, while preserving the advantages. The proposed insights and methods are formulated around the new *supervised reranking* paradigm, where supervised learning is introducing in the process of learning the reranking criteria, however, without jeopardizing the search scalability. More specifically, through supervised learning, the reranking criteria become less heuristic and better informed by the properties of the visual content in the target image collection. At the same time, the scalability is preserved by keeping the criteria *query independent*.

Two novel supervised reranking approaches are presented, namely the *feature-based* supervised reranking approach in **Chapter 2** and the *prototype-based* supervised reranking approach in **Chapter 3**. In the feature-based approach, human supervision is introduced to learn the optimal combination of reranking criteria from a set of predefined criteria. The prototype-based approach includes a more sophisticated analysis of the initial search results list before deploying it as input into the reranking process to further improve the reranking foundations. We conclude the first part of the thesis with **Chapter 4**, where we reflect upon the achievements in the domain of visual search reranking, perform a categorization and a comparative study of the methods proposed so far (including those from

this thesis) and make recommendations for future research in this direction.

1.6 Part II: Leveraging context for example-based image search

Example-based image search is usually done based on feature matching. In other words, whether an image is relevant to and returned as a result for a query is determined by the similarity between the query and the image in the visual feature space. The relevance model based on visual feature matching generally shows critical deficiencies in the image search practice due to two *gaps*: the *intent gap* [139] and the *semantic gap* [95]. The semantic gap represents the difficulty for the search engine to determine "what" the image is about or what semantic concepts it contains from its signal-level representation. The intent gap represents the difficulty to deduce the "why" behind the users search request: from all the images of "cars" returned by the search engine, which one of these best responses to the users search request? We refer to "what" and "why" aspects of the search request as the components of the users *information need* behind the query [33].

Since the feature-based image representation is given, more about the users information need could be inferred only if information sources additional to the query formulation itself are consulted. The context in which the query is formulated and the context of the images in the collection are the information sources that could prove useful for this purpose and that we investigate in this part of the thesis.

The use of contextual information for the benefit of search has already been recognized in the field of the traditional information retrieval (IR). Context-based information retrieval, which takes the context of the query and the context of document generation explicitly into the loop to better satisfy the users information need to provide better search experience, has even been recognized as a long-term challenge [9][15][12] in the IR community. Various categories of contexts, including the user profile and the texts contextualizing a keyword-based query and the terms in the collection, have been investigated and methods have been proposed to incorporate these contexts into the retrieval process [55].

With the increasing contextualization of the images on the Web, the importance of contextual information for enriching the Web image indexing and search processes has grown rapidly over the past years [24][94][32]. The main idea behind relying on the contextual information in the image search case is that an image never appears in isolation. At the image capturing stage, the metadata such as the camera exposure time, the ISO, the time when it was captured, and the GPS coordinates are often associated with the image to indicate the context in which it is captured. As described by Davis et al [24] this type of contextual information could in some cases even serve as a reliable indicator of the actual content captured by the image (e.g. a landmark). Furthermore, when an image is shared on the web, it is embedded in the web context. Different expressions of this context, e.g. a graph of hyperlinks [65] or the social network context [32], have already

been introduced as useful sources of information not only for inferring and enriching metadata, but also for revealing links between user preferences and images and in this way biasing the relevance estimation towards the "right" images in the collection. As a related example, images typically occur in the context of other images taken at the same location, uploaded by one and the same user or shared among the users. Analyzing the metadata from these images can create pointers to those metadata that can propagate from one image to another one [119], or to metadata that is more relevant to the content of the images than others (e.g. [58]). Finally, an object captured in an image is typically also not captured in isolation but in the context of a scene where other objects or scene elements can be visible as well. Deploying this scene context can help infer or confirm the metadata related to the object and verify the relevance of the captured image to the query in general.

Motivated by the great potential of the contextual information to help general image search, we focus in the second part of the thesis on the example-based image search scheme and investigate the use of various categories of contextual information for improving the robustness of this specific scheme against deficiencies of visual and textual metadata and for increasing the reliability of image search in view of the users' information need. We provide three technical contributions in this respect. We firstly focus on the main object captured in an image as the search target and explore in chapters 5 and 6 the possibilities to improve the object retrieval model by exploiting the contextual information derived from the visual scene where the object is captured from, first based on a single image serving as the query (**Chapter 5**) and then based on an image sequence (a video) taken about the target object (**Chapter 6**). The scene context proves to be helpful in the cases where the target object is either small, cluttered or occluded. While the methods presented in chapters 5 and 6 focus only on the visual scene context of the target object and therefore operate in the visual feature domain only, **Chapter 7** goes beyond this domain and proposes a unified context model that integrates different classes of contextual information related both to the query image and the images in a web collection. Inclusion of the so-called *local* and *global query context* in the image search process, as proposed in Chapter 7, was shown to significantly improve the performance of the example-based web image search compared to the related work. We conclude the second part of the thesis with **Chapter 8** where we reflect upon the achievements in deploying contextual information for improving web image search and make recommendations for future research in this direction.

1.7 How to read the thesis

The technical part of this thesis consist of original publications that have been adopted as Chapters 2-7. The references to the publications are given at the beginning of each chapter. As a consequence of working with original publications,

the notation and terminology may vary slightly across chapters. For the same reason, the introductory parts and related work sections in the chapters addressing the same general topic may be similar in terms of argumentation and the material they cover.

1.8 Full list of publications related to the thesis

1. L. Yang and A. Hanjalic. Learning to rerank web images. *IEEE Multimedia Magazine*, To Appear.[Chapter 4]

2. L. Yang, Y. Cai, A. Hanjalic, X.-S. Hua, and S. Li. Searching for images by video. *International Journal of Multimedia Information Retrieval*, pages 1–13, 2012. [Chapter 6]

3. L. Yang and A. Hanjalic. Prototype-based image search reranking. *IEEE Transactions on Multimedia*, 14(3-2):871–882, 2012. [Chapter 3]

4. L. Yang, B. Geng, A. Hanjalic, and X.-S. Hua. A unified context model for web image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 8(3):28, 2012. [Chapter 7]

5. L. Yang, Y. Cai, A. Hanjalic, X.-S. Hua, and S. Li. Video-based image retrieval. In *ACM Multimedia*, pages 1001–1004, 2011.

6. L. Yang and A. Hanjalic. Learning from search engine and human supervision for web image search. In *ACM Multimedia*, 2011.

7. L. Yang, B. Geng, Y. Cai, A. Hanjalic, and X.-S. Hua. Object retrieval using visual query context. *IEEE Transactions on Multimedia*, 13(6):1295–1307, 2011. [Chapter 5]

8. L. Yang, B. Geng, A. Hanjalic, and X.-S. Hua. Contextual image retrieval model. In *CIVR*, 2010.

9. L. Yang and A. Hanjalic. Supervised reranking for web image search. In *ACM Multimedia*, pages 183–192, 2010. [Chapter 2]

Part I

Visual Reranking for Keyword-based Image Search

Chapter 2

Supervised Reranking for Web Image Search

[1]

In this chapter we introduce the idea of *supervised rerarnking* and propose a corresponding algorithmic framework for reranking web images. Inspired by the success of the "learning-to-rank" idea proposed in the field of information retrieval, we build this framework on the "learning-to-rerank" paradigm, which derives the reranking function in a supervised fashion from the human-labeled training data. Although supervised learning is introduced, our approach does not suffer from scalability issues since a unified reranking model is learned that can be applied to all queries. In other words, a query-independent reranking model will be learned for all queries using query-dependent reranking features.

[1] This chapter was published as: Linjun Yang, Alan Hanjalic, "Supervised Reranking for Web Image Search," *Proc. ACM Multimedia 2010* [126].

Figure 2.1: *Illustration of problem cases related to text-based image search. (a) Mismatch between the image and its surrounding text. (b) Insufficient capability of the surrounding text to reveal different relevance levels of returned images for the query "George W. Bush".*

2.1 Introduction

Most of the existing Web image search engines [1, 2, 3] index images based on the associated textual information, such as the surrounding text, anchor text, URL, etc. Then the classic information retrieval (IR) techniques, which are originally designed for text retrieval, can be directly adapted for image search. Though the text-based image search approach has proven to be effective and efficient for large-scale image collections in most of the situations, it suffers from essential difficulties, which are caused mainly by the incapability of the associated text to appropriately describe the image content. For example, Fig. 2.1(a) illustrates a mismatch between the image and the surrounding text, which results in the irrelevant images being returned in the top of the result list. Fig. 2.1(b) shows some images returned based on the query "George W. Bush". Though their associated text contains the word "George W. Bush" and the images are all relevant, their relevance levels are different and this difference cannot be revealed by relying solely on the textual information.

To address the difficulties illustrated above, considerable research effort has been invested in the past years to develop the paradigm of image search using trained semantic concepts [73, 96]. There, first a model is learned per semantic concept (e.g. "tree", "George W. Bush", "car", "landscape") in a supervised fashion, and then, based on the model fit, the probability is estimated whether a given image contains a particular semantic concept. However, the recent results reported within the TRECVID evaluation benchmark [18, 97] have shown that only a limited success can be achieved using this paradigm. The first problem lies in the insufficient capability of the paradigm to scale up to a large number of concepts that are required to cover a realistic query space [34]. The second problem lies in the semantic gap between the abstractness of the semantic concepts and the low-level image features used to train the concept models.

As an alternative to the search paradigm described above, *visual search reranking*, has attracted increasing attention from both academia and industry [37, 38, 43, 102, 121]. Generally speaking, visual search reranking is devised to re-order the image search result list returned by the text search engine by exploiting the visual information contained in the images. While considering an additional (visual) modality is expected to make this new search paradigm more powerful than the pure text-based search, this paradigm also scales better than the one based on semantic concepts since it does not require offline model learning.

The scalability advantage of the visual search reranking paradigm stems from its *unsupervised* nature, i.e., from the unsupervised approach to learning the reranking function that is used to refine the initial search result. However, this approach also makes it difficult to handle some of the key problems encountered in the image search practice. The reranked image search result is typically based on heuristic assumptions regarding the role of the visual modality in determining the relevance of the search results and the relative importance of the visual modality compared to the initial text-based search result that it is supposed to "correct". Since these assumptions may not be valid to the same degree in different use cases (search engines), the reranking performance remains largely unpredictable.

In this chapter we build on the basic visual search reranking idea and address its deficiencies specified above by introducing a supervision step into the reranking process. Through this step, the possibility is created to employ information from within the data collection to steer the reranking process and to reduce the need for making heuristic assumptions. We refer to this further as the *supervised reranking* or *learning-to-rerank* paradigm.

Compared to the classic supervised approaches to multimedia search related to semantic concept learning, the scalability of our approach is not degraded by introducing supervision. This is because the ground truth information available for only a limited number of queries is used to learn a *generic* reranking model to handle all queries. In other words, different from semantic concept learning, which learns *query-dependent* models using *query-independent* features , learning-to-rerank embeds the query information into the *query-dependent* reranking features, which estimate the relevance between the query and an image in the collection. Then a *query-independent* model is learned and employed to rerank images for all queries. This decomposition of query-dependency into the reranking features in the learning-to-rerank paradigm also makes the reranking function better "learnable" than an arbitrary semantic concept, imposes fewer requirements on the training data set and requires less manual annotation effort than in the case of semantic concept learning. Moreover, the implicit user feedback (e.g. click-through log), can also be employed as a source of training data for this purpose [44] The scheme of our approach illustrating the offline step of learning a general reranking model from the labeled data and then applying the learned model online to handle all queries in a given search use case is illustrated in Fig. 2.2.

After we position our proposed reranking approach in Section 2.2 with respect

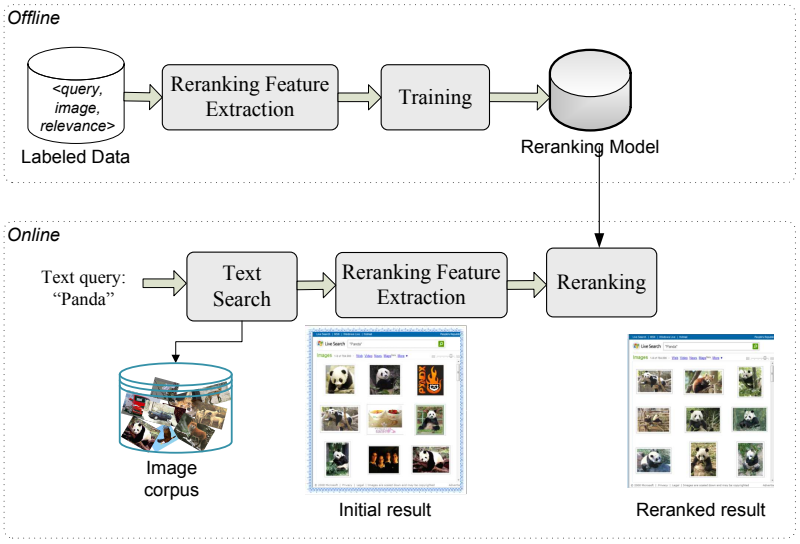

Figure 2.2: *The overview of feature-based supervised reranking approach.*

to the related previous work, we formulate the learning-to-rerank paradigm in Section 2.3 and explain the issues related to its practical implementation. This is followed in Section 2.4 by introducing the features that we compute from the collection and use as input into the reranking mechanism. In Section 2.5 we first describe the experimental setup we devised to evaluate the performance of the developed learning-to-rerank mechanism. Then we present the results of the evaluation at various levels and provide a discussion regarding the effectiveness of the design choices we made when developing the mechanism. A list of suggestions for future work in this direction is provided in Section 2.6. A brief conclusion in Section 2.7 completes the paper.

2.2 Related Work

The existing visual search reranking methods are all unsupervised and can be defined as the classification-based [121], graph-based [102] and clustering-based [37] methods.

Classification-based methods [121, 60, 132, 88] first select some pseudo-relevant samples from the initial search result. Then a classifier or a ranking model is learned with the pseudo-relevant/-irrelevant samples serving as training data. The classification output scores for each image are then used to generate the final ranking. Such methods are also referred to as pseudo-supervised methods, since they rely on the initial ranking to automatically acquire training data for the purpose of learning a query-dependent ranking model. In the clustering-based

methods [37], the images in the initial result are firstly grouped automatically into several clusters. Then the reranked result list is created first by ordering the clusters according to the cluster conditional probability and then by ordering the samples within a cluster based on their cluster membership value. The more recently proposed graph-based methods [38, 43, 102] were demonstrated to be more effective for Web-scale image search and have therefore received increased attention. Firstly, a graph is built with the images in the initial result serving as nodes. An edge is defined between two images if they are visual neighbors of each other and the edges are weighted by the visual similarities between the images. Then, reranking can be formulated, for instance, either as a random walk over the graph [38, 43] or an energy minimization problem [102].

Learning-to-rerank has a similar underlying rationale as the learning-to-rank paradigm that is known from IR [17, 44] and that was shown to be superior to the existing classical unsupervised IR methods, such as Okapi BM25 [82] and *tf-idf* [67]. While both paradigms utilize the human labeled data to make a ranking/reranking model better fit the application scenario and human expectation of what the search result should be, learning-to-rerank has several unique characteristics. First, in learning-to-rerank the initial ranking result from the text-based search serves as a prior, which needs to be effectively incorporated into the reranking process. Second, in learning-to-rerank the query and the documents have different representations, i.e., the query is textual while an image is visual. This poses a considerable challenge on the design of reranking features that we address in this chapter through a careful feature engineering step.

2.3 Learning to Rerank

In order to elegantly incorporate the supervised learning step into the reranking approach we present in this chapter a general formulation of the learning-to-rerank problem and decompose it into two key components: the learning step and the feature design step. For the learning step, the Ranking SVM adopted from the learning-to-rank approach is adjusted to solve the learning problem in the new reranking context. For the feature design, motivated by the existing successful reranking methods, we design an 11-dimensional vector of reranking features based on the exploitation of the visual context, initial ranking, and the pseudo relevance feedback. The overview of the learning-to-rerank system and the constituent components is illustrated in Fig. 2.2.

2.3.1 Formulation

We formulate the problem addressed in this chapter through the definitions given below.

Definition 2.3.1. *A ranking* $\mathbf{r}(\mathcal{D})$*, abbreviated as* $\mathbf{r}$*, is a function mapping the document set* $\mathcal{D}$ *to the vector of documents' rankings. In other words, each element*

of $\mathbf{r}$*, defined as* $r(d_j)$ *and further abbreviated as* r_j*, is the ranked position of the document* $d_j \in \mathcal{D}$*.*

Specifically for the image search context, $\mathcal{D}$ will be used to denote the set of image documents returned by the initial text search.

Definition 2.3.2. *A reranking model is defined as a function*

$$\mathbf{r} = f(\mathcal{D}, \bar{\mathbf{r}}, q), \tag{2.1}$$

where $\bar{\mathbf{r}}$ *is the ranking of documents in the initial search result, and* q *is the query.*

Generally, the reranking aims at returning a new ranked list after taking the query, the initial result and initial ranking as input. Usually, the query term q can be ignored from Eqn. (2.1) since it has been reflected in the initial result $\mathcal{D}$ and ranking $\bar{\mathbf{r}}$. In view of the above, we can define the objective of the learning-to-rerank approach as to learn the reranking model f from training data.

Definition 2.3.3. *Learning-to-rerank is defined as a process of learning a reranking model* f *from the given training samples* $\{\mathcal{D}^i, \bar{\mathbf{r}}^i, q^i, \tilde{\mathbf{r}}^i\}$*, where* $\mathcal{D}^i, \bar{\mathbf{r}}^i, \tilde{\mathbf{r}}^i$ *are the initially returned documents, initial ranking and the ground-truth ranking corresponding to the query* q^i*. The learning process can be formulated as the process of minimizing the loss function*

$$f^* = \arg\min_f \sum\nolimits_i \triangle(\tilde{\mathbf{r}}^i, f(\mathcal{D}^i, \bar{\mathbf{r}}^i, q^i)), \tag{2.2}$$

where $\triangle$ *measures the loss between the ground-truth* $\tilde{\mathbf{r}}^i$ *and the prediction* $\hat{\mathbf{r}}^i = f(\mathcal{D}^i, \bar{\mathbf{r}}^i, q^i)$*.*

Since the definition given above is rather general, it leaves several issues to be addressed more explicitly in order to be able to realize the cost minimization in Eqn. (2.2) in a practical case. First, a function definition of the reranking model f is needed. Second, the learning algorithm should be specified. Finally, the loss function needs to be designed carefully and the optimization problem should be solved efficiently. In the following subsections, we will introduce our approach to addressing these open issues.

2.3.2 Reranking Model

It is commonly recognized that in visual search reranking there are two cues which can be taken into account to obtain a refined ranked list [102]. One is the initial ranking obtained from text-based search, which often shows acceptable ranking performance though is often affected by noise due to the imperfect match between the surrounding text and the image's content. The other one is the visual content of the ranked documents, which can be regarded as the visual context in which

user's information need is formulated. Based on such an analysis we can design the reranking model by combining the two cues as follows,

$$f(\mathcal{D}^i, \bar{\mathbf{r}}^i, q^i) = \arg\max_{\mathbf{r}^i} -w_0 D(\mathbf{r}^i, \bar{\mathbf{r}}^i) + \sum\nolimits_{k=1}^{Z} w_k \mathbf{r}^i \cdot [\psi_k(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)]_{j=1}^{P}. \quad (2.3)$$

Here d_j^i is the j-th image in the sorted list of images in the initial result $\mathcal{D}^i$ based on the search engine. D is the distance between two rankings. ψ_k is a function steering the reranking feature extraction processes applied to an image d_j^i in the initial ranking result and P is the number of images in the initial result to be considered in reranking. Z is the number of predefined reranking features and $\mathbf{w} = [w_0, w_1, \cdots, w_Z]^T$ are the corresponding weighting coefficients. The basic idea of Eqn. (2.3) is to maximize the bias of the reranked list towards the initial one while at the same time maximizing the coherence of the images ranked similarly in terms of the reranking features.

There are different methods to measure the distance between two ranking lists, such as the Normalized Discounted Cumulative Gain (NDCG) [41] and Kendall's τ ranking correlation [46]. However, by incorporating such distances into Eqn. (2.3), it would be difficult to obtain a closed-form solution, which would make the learning of $\mathbf{w}$ much more difficult. Besides, it would be convenient if the solution of the problem (2.3) is a linear function w.r.t. $\mathbf{w}$, so that the resulting learning problem could be solved easily and the online ranking process could be more efficient. Under such guidelines, we propose to compute the ranking distance by transforming the initial ranking into a score vector,

$$D(\mathbf{r}^i, \bar{\mathbf{r}}^i) = -\mathbf{r}^i \cdot \mathbf{s}(\bar{\mathbf{r}}^i), \quad (2.4)$$

where $\mathbf{s}(\bar{\mathbf{r}}^i)$ is the score vector with $s(\bar{r}_j^i)$ corresponding to the ranking score of the image ranked at the j-th position in the initial ranked list.

By substituting Eqn. (2.4) into Eqn. (2.3), the reranking model can be formulated as,

$$f(\mathcal{D}^i, \bar{\mathbf{r}}^i, q^i) = \arg\max_{\mathbf{r}^i} \sum\nolimits_{k=0}^{Z} w_k \mathbf{r}^i \cdot [\psi_k(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)]_{j=0}^{P}, \quad (2.5)$$

where $[\psi_0(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)]_{j=0}^{P} = \mathbf{s}(\bar{\mathbf{r}}^i)$. It is easily derived that the solution is the ranking of images according to the score vector $\sum_{j=0}^{Z} w_k [\psi_k(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)]_{j=0}^{P}$.

Consequently we achieve a linear model for reranking by combining different reranking features, where the initial ranking is also represented as one of the features. The model is similar to the ranking function widely used in the learning-to-rank approach. Hence, the classic learning-to-rank algorithm, such as Ranking SVM [44], could be adopted for the learning-to-rerank paradigm as well. In the following section, we will introduce the standard algorithm of Ranking SVM and a modification we introduced in this algorithm to adapt it to our reranking problem.

2.3.3 Adaptation of Ranking SVM

Technically speaking, the objective of the learning-to-rerank task is to estimate the parameters by minimizing a loss function. Methods that can be used for this purpose differ in the design of the loss function. Ranking SVM [44] is a classic algorithm applied in learning-to-rank, where the loss function is defined as a combination of the prediction loss and the regularization term:

$$\begin{aligned} &\triangle(\mathbf{r}^i, f(\mathcal{D}^i, \bar{\mathbf{r}}^i, q^i)) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \\ &C\sum\nolimits_{d_j^i \succ_{\mathbf{r}^i} d_k^i} \max(0, 1 - \mathbf{w}^T(\Psi(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i) - \Psi(d_k^i, \mathcal{D}^i, \bar{\mathbf{r}}^i))), \end{aligned} \tag{2.6}$$

where the first term is the regularization and the second one is the hinge loss on the document pairs. Here, $d_j^i \succ_{r^i} d_k^i$ denotes that the image d_j^i is ranked before the image d_k^i in the ranked list $\mathbf{r}^i$. C is the trade-off parameter. $\Psi(d_k^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)$ is the reranking feature vector of d_k^i.

By substituting the loss Eqn. (2.6) into the problem (2.2), we obtain the following optimization problem:

$$\begin{aligned} \min\ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum \xi_{jk}^i \\ s.t.\ & \forall i, d_j^i \succ_{\mathbf{r}^i} d_k^i : \mathbf{w}^T(\Psi(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i) - \Psi(d_k^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)) \geq 1 - \xi_{jk}^i \\ & \forall i, j, k : \xi_{jk}^i \geq 0, \end{aligned} \tag{2.7}$$

where ξ_{jk}^i is the slack variable.

We can clearly see from (2.7) that the rationale behind the Ranking SVM is that it models the prediction loss based on the preference between two documents. Then, the learning-to-rank problem can be reduced to the classification of the preference over document pairs.

It is important to note, however, that in the reranking problem, the features are of different importance. First, while Z dimensions attribute to the visual content analysis, only one dimension is related to the initial ranking. Moreover, the initial ranking is an important information source for reranking since it often gives a reasonable result. Since in the problem formulation (2.7) the influence of the initial ranking is likely to be degraded, and even severely degraded if Z is large, we modify the problem (2.7) to allow the initial ranking to provide a larger contribution, if necessary. The modified optimization problem can be formulated as

$$\begin{aligned} \min\ & \frac{1}{2}((\frac{w_0}{\alpha})^2 + \sum\nolimits_{t=1}^{Z} w_t^2) + C\sum \xi_{jk}^i \\ s.t.\ & \forall i, d_j^i \succ_{\mathbf{r}^i} d_k^i : \mathbf{w}^T(\Psi(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i) - \Psi(d_k^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)) \geq 1 - \xi_{jk}^i \\ & \forall i, j, k : \xi_{jk}^i \geq 0. \end{aligned} \tag{2.8}$$

where α is the parameter to control the confidence of the corresponding feature from the initial ranking. We empirically set it to be equal to Z.

Approaches to solving the standard classification SVM, such as SMO (Sequential Minimal Optimization) [78], can be used directly to solve the problem (2.8). In this chapter, we adopt the fast algorithm based on the cutting-plane method [45].

2.4 Features

In this chapter we envision three types of reranking features. Considering the result of the initial text search as a strong prior, the features $\psi_0(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)$ based on the initial ranking are referred to as the *prior reranking features*. Other feature types are content based and extracted through visual content analysis. In this chapter, we propose to extract the content-based reranking features from two perspectives. The first leads to the *contextual reranking features*, which are extracted by considering the images from the initial ranking result as a visual context of the target image. The second leads to the *pseudo relevance feedback features*, which are extracted by considering the top N images as positive examples and then by ranking the others based on these examples. The three approaches lead to 11 reranking features being extracted, corresponding to $\psi_k(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i)$ for $k = 1...11$, which are summarized in Table. 2.1.

2.4.1 Contextual Reranking Features

Visual context of an image in the initial result reflecting the neighborhood structure of the ranked item list is a useful information source to refine the initial search result list, as shown by numerous existing methods for visual search reranking [102][38]. In this section, we present a simple yet effective method to exploit the visual context information for image search reranking.

Visual context

Given an image $d_j \in \mathcal{D}$, where $\mathcal{D}$ is the initial result returned by the text search for query q, its visual neighbors in $\mathcal{D}$ can be computed based on the following three strategies.

- *K-Nearest Neighbors (KNN)*: The top K similar images in $\mathcal{D}$ to d_j are regarded as the neighbors of image d_j.
- *ϵ-Nearest Neighbors (ϵNN)*: The images in $\mathcal{D}$ whose distance to d_j is less than the threshold ϵ are regarded as the neighbors of image d_j.
- *The combination*: The images which satisfy the above two strategies simultaneously are regarded as the neighbors of image d_j in $\mathcal{D}$.

The neighbors obtained above are sorted according to the visual similarity to d_j to get the list of images $\mathbf{N}^j$. N_k^j stands for the *k-th* image on the ranked neighbors list.

Recently a new neighborhood structure, the so-called *reciprocal neighborhood* [49], has been proposed and successfully applied in image search reranking [107]. Basically, if an image d_i occurs as the neighbor of image d_j then d_j is referred to as a reciprocal neighbor of d_i. Formally, given an image d_i, the set of its reciprocal neighbors can be defined as

$$\mathcal{R}^i = \{d_j | d_i \in \mathbf{N}^j\}, \tag{2.9}$$

and the reciprocal neighbor list $\mathbf{R}^i$ is the sorted list of images in $\mathcal{R}^i$ according to the visual similarity between d_i and d_j.

The visual context of an image is defined by its neighbors and reciprocal neighbors taken from the initial search result. Based on this, we now proceed with the extraction of the contextual reranking features, as explained in the next subsections.

Neighborhood rank voting

A straightforward approach to utilizing the visual context information for extracting the reranking features is neighborhood voting. There are different variants of the neighborhood voting. In *hard voting*, each of the neighbors contributes equally to the relevance of the image d_j, that is

$$HV_N(d_j) = len(\mathbf{N}^j), \tag{2.10}$$

where $len(\mathbf{N}^j)$ is the size of the vector $\mathbf{N}^j$.

It can easily be observed that the hard voting score corresponds to the set cardinality of a neighborhood. Therefore, hard voting is effective only in case of applying the ϵ-Nearest Neighbor strategy or the combination strategy in the construction of the visual context. A drawback of hard voting is that all neighbors are treated equally. Different neighbors should namely contribute differently to the relevance of a target image according to their own relevance, which can be expressed through their initial ranking or their position in the ranked neighborhood. We refer to such more sophisticated voting as *soft voting*. Soft voting based on the initial ranking assigns weights to the votes using the following expression:

$$RSV_N(d_j) = \sum_{k=1}^{len(\mathbf{N}^j)} \frac{1}{\log(\bar{\mathbf{r}}(N_k^j) + 1)}. \tag{2.11}$$

The transformation from the initial ranking to the voting score using the log function is motivated by the discount term in NDCG (Normalized Discounted Cumulative Gain) [41], which assigns a larger relative importance to top images in the returned result since their relative relevance to the query is assumed larger.

Furthermore, the voting score of each neighboring image can be weighted by its adjacency to the target image, measured by its position in the ranked

neighborhood. Hence, the soft voting based on a neighbor-rank-weighted initial ranking can be computed using the following expression,

$$NRSV_N(d_j) = \sum_{k=1}^{len(\mathbf{N}^j)} \frac{1}{\log(\bar{\mathbf{r}}(N_k^j) + 1) \times k}. \tag{2.12}$$

Reciprocal neighborhood rank voting

Similar to the neighborhood rank voting, the reciprocal neighborhood rank voting can also be divided into hard and soft voting. The corresponding reranking features can be computed using the analogy to Eqn. (2.10, 2.11, 2.12) as

$$HV_R(d_j) = len(\mathbf{R}^j). \tag{2.13}$$

$$RSV_R(d_j) = \sum_{k=1}^{len(\mathbf{R}^j)} \frac{1}{\log(\bar{\mathbf{r}}(R_k^j) + 1)}, \tag{2.14}$$

$$NRSV_R(d_j) = \sum_{k=1}^{len(\mathbf{R}^j)} \frac{1}{\log(\bar{\mathbf{r}}(R_k^j) + 1) \times NR(R_k^j, d_j)}, \tag{2.15}$$

where $NR(R_k^j, d_j)$ is the ranked position of d_j among the neighbors of image R_k^j.

In addition to the features mentioned above, the reciprocal neighborhood also has some unique characteristics which can be exploited for reranking feature extraction. In particular, we focus here on the ranked position of the target image in the neighborhoods of the reciprocal neighboring images, which represents how confidently other images select the target image as a neighbor. Hence, we define the soft voting, which takes only the reciprocal neighborhood rank into consideration as

$$NSV_R(d_j) = \sum_{k=1}^{len(\mathbf{R}^j)} \frac{1}{NR(R_k^j, d_j)}. \tag{2.16}$$

2.4.2 Pseudo relevance feedback

Pseudo relevance feedback (PRF) is a technique widely used in information retrieval and recently also adopted in solving the visual search reranking problem [121, 60, 132]. The basic idea of PRF in the visual search reranking context is to regard the top ranked images in the initial result as the relevant ones, and then to apply a relevance feedback technique on this "pseudo" relevant image set to refine the search result. Although the true relevance of the top-ranked images is unknown since human is left out of the loop, the results shown in Fig. 2.3, which are obtained on a representative image collection, using human judgment as a reference, indicate that the top m images in the initial text-based search could be considered more relevant than the lower-ranked ones. In the following, we will elaborate on three light-weight PRF approaches to compute the reranking features.

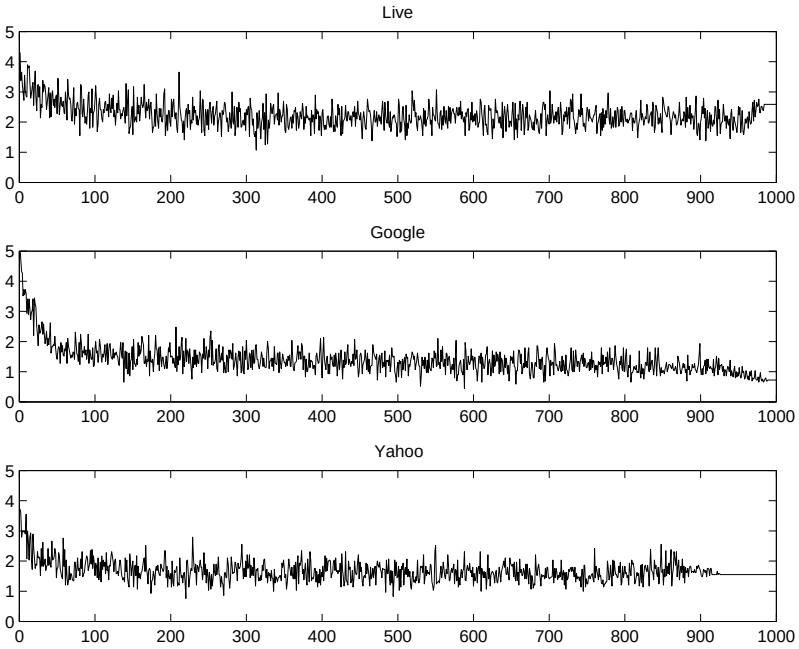

Figure 2.3: *The distribution of images' relevance (judged by human oracle) at different positions of the initial ranking. The horizontal axis is the ranking position and the vertical axis is the mean relevance score among all the queries. The dataset used for the study is detailed in Section 2.5.1.*

Given the top m pseudo relevant images, a reranking feature of d_j can be computed by estimating its relevance score as the probability of generating d_j from these m images,

$$PRF(d_j) = p(d_j|d_1, d_2, \cdots, d_m), \tag{2.17}$$

where p is the probability density function and $d_1, d_2, \cdots, d_m$ are the top m images in the initial result. A typical approach to estimating the function p is the kernel density estimation (KDE). The KDE based PRF feature is extracted as follows,

$$PRF_d(d_j) = \frac{1}{m} \sum\nolimits_{k=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \delta(d_j, d_k), \tag{2.18}$$

$$\delta(d_j, d_k) = exp(||d_j - d_k||^2/(2\sigma^2)), \tag{2.19}$$

where δ is the RBF (radial basis function) kernel and σ represents the standard deviation. Here $||d_j - d_k||^2$ is the Euclidean distance of two images d_j and d_k based on visual features.

Alternative methods are based on duplicate voting and therefore conceptually similar to the approach we used to compute the contextual features. However, while in the computation of the contextual features we use the entire initial result list to construct the visual context, in PRF we estimate the relevance of the target images based solely on the m top-ranked images.

A PRF based on hard and soft duplicate voting can be defined using the formulas in Eqn. (2.20) and Eqn. (2.21), respectively:

$$PRF_{dv}(d_j) = \frac{1}{m} \sum\nolimits_{k=1}^{m} IsDup(d_j, d_k), \tag{2.20}$$

$$PRF_{sdv}(d_j) = \frac{1}{m} \sum\nolimits_{k=1}^{m} \frac{IsDup(d_j, d_k)}{\log(\bar{\mathbf{r}}(d_k) + 1)}, \tag{2.21}$$

where the function $IsDup$ can be any duplicate detection function [77]. In the experiments reported in this chapter, we simply use a threshold to determine whether two images are duplicates based on their visual similarity. Information about the features used for visual similarity computation is provided in Section 2.5.1.

We note here that PRF-based reranking features may lead to a degradation of the performance compared to the initial ranking due to the potential query drifting problem. However, in the approach introduced in this chapter, the weights applied to the PRF-based ranking features are learned from the human labeled ranking. Therefore, if the PRF feature does not perform well in a given use case, it will receive a low weight and will not influence the final result to a large extent. Hence the query drifting problem, if present, can be alleviated in this way.

2.4.3 Initial ranking

As stated before, the initial ranking provides critical input information for the reranking step. In most cases, only the ranking position is available, and not the ranking scores from a search engine. Working with the initial ranking position directly as a feature would not be a good option since the Web search ranking is normally optimized for the top results. For example, the widely used evaluation measure NDCG, which is also the optimization objective of typical learning-to-rank methods [20], employs a discount factor to emphasize the top-ranked items. In other words, the top images in the text-based search result have larger confidence to be relevant and the confidence degrades super-linearly with the increasing ranking position. Hence, we choose to transform the initial ranking position by following an analogy to the discount factor in NDCG to obtain the following feature which still reflects the initial ranking, but also takes into account the non-linearity of the relevance confidence degradation:

$$IR(d_j) = 1/\log(j + 1), \tag{2.22}$$

where j is the position of the image d_j in the initial ranked list.

Table 2.1: *An overview of the proposed reranking features.*

IR	Initial Ranking
HV_N	Hard Voting of Neighbors
RSV_N	Initial Rank based Soft Voting of Neighbors
$NRSV_N$	Neighbor Rank Weighted Initial Rank based Soft Voting of Neighbors
HV_R	Hard Voting of Reciprocal Neighbors
RSV_R	Initial Rank based Soft Voting of Reciprocal Neighbors
NSV_R	Neighbor Rank based Soft Voting of Reciprocal Neighbors
$NRSV_R$	Neighbor Rank Weighted Initial Rank based Soft Voting of Reciprocal Neighbors
PRF_d	Local Density Estimation for PRF
PRF_{dv}	Duplicate Voting for PRF
PRF_{sdv}	Soft Duplicate Voting for PRF

2.5 Experiments

In this section we first describe the experimental setup we used to evaluate our proposed learning-to-rerank paradigm. Then we present the results of the evaluation at various levels and provide a discussion regarding the effectiveness of critical design choices.

2.5.1 Experimental setup

We conduct the experiments on two datasets: a collected Web image dataset and a publicly available MSRA-MM dataset [114]. We explain both datasets in the following paragraphs.

29*3 queries Dataset: The dataset we used for the experiments reported in this chapter consists of 73,340 images collected from three most popular commercial image search engines, i.e., Google, Live and Yahoo. We selected 29 queries from the query log of a commercial image search engine and popular tags of Flickr. These queries cover a vast range of topics, such as scenes (*Sky* and *Winter*), objects (*Grape* and *Panda*) and named person entities (*George W. Bush*). The queries are listed in Table 2.2. For each query, at most top 1000 images returned by each of the three search engines are collected.

For each image, its relevance degree with respect to the corresponding query is judged by three human judges and using four relevance levels, i.e., "Excellent", "Good", "Fair" and "Irrelevant". Then, for each image, the final ground truth relevance is defined as the median of the scores given by the three judges.

To analyze the images' visual content and compute the distance between images, we adopt 7 widely used low-level visual features to represent the image:

Table 2.2: *The queries in the 29*3 queries dataset.*

Animal, Beach, Beijing Olympic 2008, Building, Car, Cat, Clouds, Earth, Flower, Fox, Funny dog, George W. Bush, Grape, Hearts, Hello Kitty, Hiking, Mercedes logo, Panda, Sky, Statue of Liberty, Sun, Trees, Wedding, White Cat, White House, White House Night, Winter, Yellow Rose, Zebra

Attention Guided Color Signature, Color Fingerprint, Multi-Layer Rotation Invariant EOH, Histogram of Gradients, Daubechies Wavelet, Facial Features, and Black & White [22].

MSRA-MM Dataset: To more comprehensively evaluate the proposed approach, we also conduct experiments on a publicly available dataset[2] which includes 68 popular queries. The detailed information about this dataset can be found at [114]. For computing the distance between images in this case, we used the features provided by the dataset in order to make our results reproducible and to enable comparisons with other approaches in the future.

For each dataset we uniformly split the queries into five folds. When evaluating each of the folds the remaining four folds are used as training. To evaluate the ranking performance, NDCG is adopted, which is a measure commonly used in information retrieval, especially when there are more than two relevance levels.

In the following we will mainly use the 29*3 queries Dataset to evaluate the proposed approach since it is diverse and includes different categories of queries and three mainstream search engines. If not explicitly stated, the experimental results reported below refer to the 29*3 queries Dataset. The MSRA-MM dataset will be used solely to demonstrate the transferability of the proposed paradigm across collections.

2.5.2 General performance evaluation

We compare the proposed learning-to-rerank method with Bayesian reranking [102, 103], random walk [38], and the text-based search baseline to demonstrate its effectiveness. The overall performance (The NDCG averaged over all queries) is shown in Fig. 2.4. Since the LocalPair variant of Bayesian reranking [103] performs the best among the seven graph-based reranking methods including six Bayesian variants and random walk, we will use it as the representative of Bayesian and graph-based reranking methods. From this result, we can see that the proposed learning-to-rerank (letorr) method consistently outperforms the text-based search baseline and Bayesian reranking at all truncation levels. Specifically, letorr obtains about 11.6% and 7.4% relative improvements on NDCG@100

[2] http://research.microsoft.com/en-us/um/people/xshua/imm2009/dataset.html

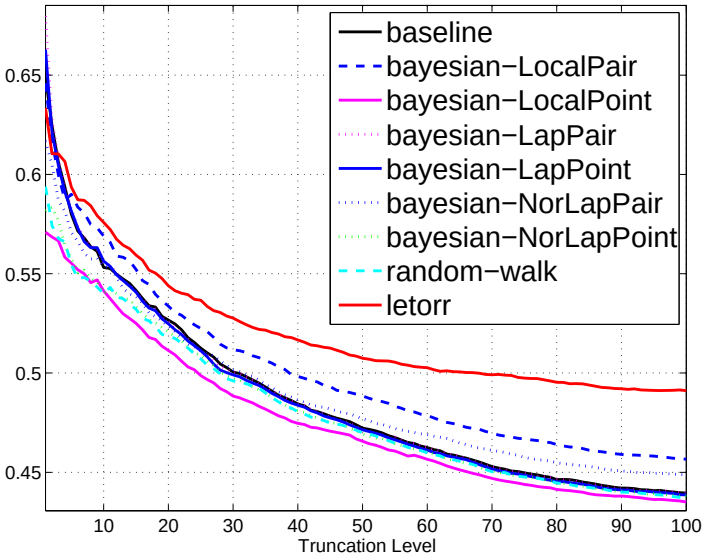

Figure 2.4: *Performance comparison between the learning-to-rerank (letorr) and seven graph-based reranking methods on the 29*3 queries dataset. The six variants of Bayesian reranking are based on three consistencies: Local learning (Local), Laplacian (Lap), and Normalized Laplacian (NorLap) and two kinds of ranking distances: Point-wise (Point) and Pair-wise (Pair). The vertical axis is NDCG.*

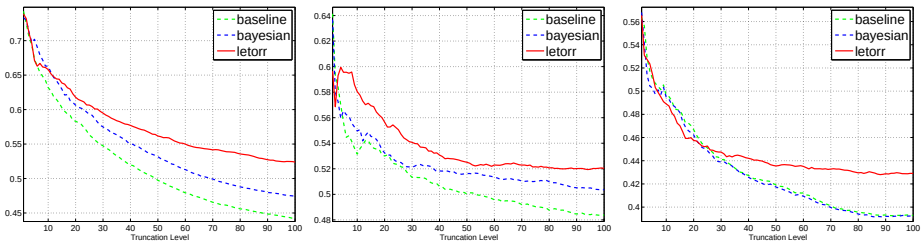

Figure 2.5: *The performance on the three search engines. From left to right: Google, Live, Yahoo. The vertical axis is NDCG.*

compared with the text-based search baseline and Bayesian reranking, respectively. In comparison, Bayesian reranking only gains 3.8% relative improvement over the text-based search baseline. The results indicate that the proposed supervised learning-to-rerank method can learn a good reranking function and that the learned reranking model can be generalized well across a broad range of queries. Moreover, the proposed method is a lightweight one since it requires less computational cost than graph-based reranking methods including the Bayesian rerank-

Table 2.3: *The p values of the significance test.*

	Letorr vs. Baseline	Letorr vs. Bayesian
NDCG@40	0.0017	0.0266
NDCG@100	5.6281e-8	4.3593e-6

ing, which requires iterative computation. The 11.6% performance improvement with less computational cost demonstrates that the learning-to-rerank method is a promising paradigm for Web image search. This conclusion is also supported by the result reported in Fig. 2.5 that shows the performance comparison on all three search engines: Live, Yahoo, and Google. We can see that for all the three search engines the learning-to-rerank method improves the performance over the text-based search baseline. Moreover, the proposed method performs consistently better than Bayesian reranking on all three search engines.

Figure 2.6 gives an example result for the illustration of the advantages of supervised learning-to-rerank over Bayesian reranking. It can be observed that the learning-to-rerank method promotes highly relevant images (the images marked by the red rectangle), which are ranked at the bottom in the initial list, to the top. We explain the inability of the Bayesian reranking to do the same by the following reasons. First, Bayesian reranking relies more on the initial ranking, while in learning-to-rerank, the initial ranking is regarded as one of many features, the weights of which can be learned and adjusted to the use cases automatically. Second, in computing the reranking features, we mainly use the visual neighborhood structure instead of the visual similarity itself. This alleviates the problems introduced by the imperfection of the visual similarity estimation that tend to make the existing reranking methods unpredictable in many practical use cases. For instance, the promotion of the grassland image (the image marked by blue rectangle) by Bayesian reranking should attribute to the deficiency of visual similarity estimation.

We further performed a statistical significance test to verify whether the improvement of the learning-to-rerank method is statistically significant. The p values of the t-test of learning-to-rerank over the text-based search baseline as well as the Bayesian reranking method in terms of NDCG@40 and NDCG@100 are shown in Table 2.3. They are computed by modeling the performance improvement for each query (NDCG difference between two methods) as a t-test. From this result we can see that the improvement of the learning-to-rerank method is statistically significant.

The performance comparison on the MSRA-MM dataset is reported in Fig. 2.7. We can see that on this dataset the performance is also greatly boosted after applying the proposed learning-to-rerank method to rerank the image search results. The learning-to-rerank method improves the performance over baseline by 8.2% and over Bayesian reranking by 4.9% in terms of NDCG@100. Due to the space limits, we will further focus only on the performance analysis on the

Figure 2.6: *An illustration of the reranking results for the query "George W. Bush" on Yahoo image search engine. The images with red rectangles are examples of highly-relevant result, while the images with blue rectangles are irrelevant.*

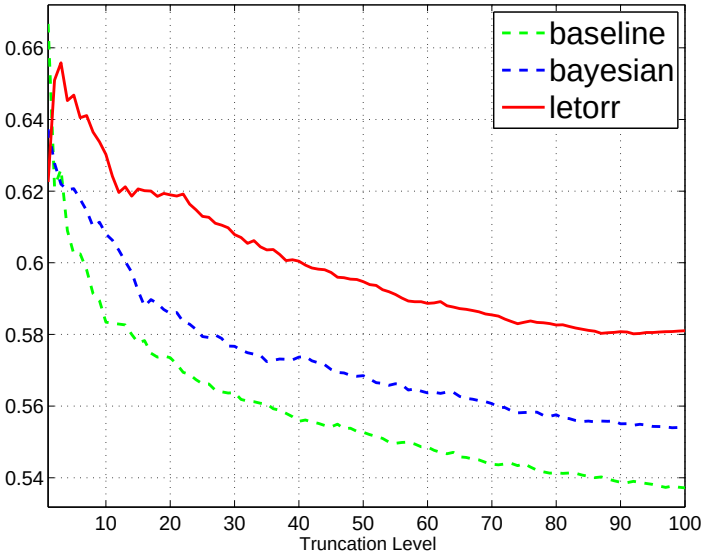

Figure 2.7: *Performance comparison between the learning-to-rerank (letorr) and Bayesian reranking on MSRA-MM dataset. The vertical axis is NDCG.*

29*3 queries Dataset.

2.5.3 Performance analysis over different queries

The performance of the learning-to-rerank method on different queries is shown in Fig. 2.8. Each bar corresponds to a combination of a query keyword and a search

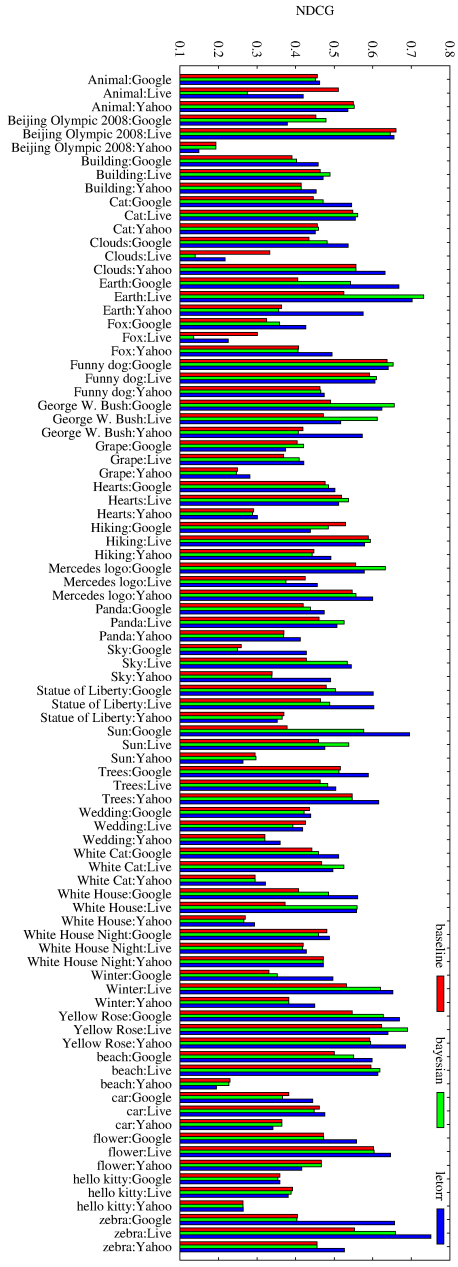

Figure 2.8: *The performance for each of the queries on the three search engines.*

Table 2.4: *The performance for each of the individual features.*

Feature	HV_N	RSV_N	$NRSV_N$	HV_R	RSV_R	NSV_R
NDCG@40	0.484	0.445	0.458	0.363	0.366	0.360
NDCG@100	0.439	0.441	0.451	0.376	0.381	0.373

Feature	$NRSV_R$	PRF_d	PRF_{dv}	PRF_{sdv}	IR	Baseline	letorr
NDCG@40	0.369	0.355	0.493	0.500	0.484	0.484	0.517
NDCG@100	0.382	0.364	0.445	0.449	0.440	0.440	0.491

engine. Therefore there are totally 87 (29*3) bars. We can see from the figure that for a majority of queries we can achieve performance boosting after applying visual search reranking. Moreover, on a large proportion of the queries the proposed learning-to-rerank method outperforms the Bayesian reranking method significantly.

The reranking performance of an individual query is related to the characteristics of the initial text-based search result. We can draw the conclusion that the queries for which the relevant images in the initial result are semantically or visually coherent, will benefit more from reranking. For example, among the 87 queries, "sun" on Google obtains the highest performance improvement after applying the learned reranking model since the relevant images are visually coherent. However, for the ambiguous queries, such as "animal", visual search reranking cannot bring out too much performance improvement. The performance is even degraded in some cases. For example, for the query "animal" on Live and Google, letorr introduces 0.09 and 0.01 performance degradation, respectively. For "animal" on Live, Bayesian reranking even degrades the performance 54% relative to the initial ranking.

As shown in Fig. 2.9, for a lot of queries the performance improvement of Bayesian reranking is near zero. This means that Bayesian reranking cannot change the initial ranking for these queries based on visual consistency. These queries can be regarded as "difficult" queries for visual reranking. We can see that nearly all the queries for which the initial ranking is poor are such "difficult" queries. Thanks to the neighborhood voting based features, which alleviate the imperfection of the visual similarity computation and the learning strategy, which makes the reranking model more adapted to the data, learning-to-rerank method can still improve the performance for these "difficult" queries.

2.5.4 Feature analysis

In this section, we will analyze the effects of the 11 selected reranking features. Each of the reranking features can be used to rank the images on their own. The ranking performance using the 11 features individually is summarized in Table 2.4. It can be clearly observed that most of the content based reranking features perform well, some of which are even better than the initial ranking. For example, the performance of PRF_{sdv} is 0.5 in terms of NDCG@40, better than 0.484 for the

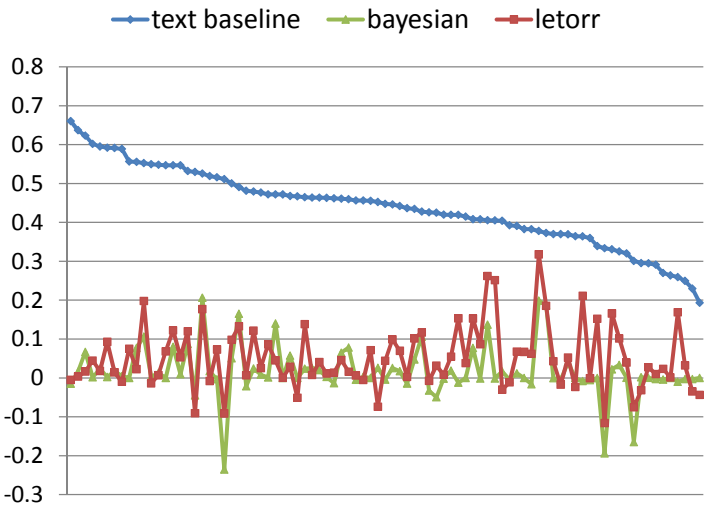

Figure 2.9: *The performance improvement (in terms of NDCG) of visual search reranking over different queries. The horizontal axis corresponds to the queries sorted in the descending order of the performance of the text baseline. For text baseline the ranking performance is plotted, while for letorr and Bayesian the performance improvement is plotted.*

Table 2.5: *The performance of learning-to-rerank by leaving one feature out.*

Feature	HV_N	RSV_N	$NRSV_N$	HV_R	RSV_R	NSV_R	$NRSV_R$
NDCG@40	0.516	0.517	0.516	0.515	0.517	0.519	0.517
NDCG@100	0.490	0.491	0.486	0.487	0.489	0.492	0.491

Feature	PRF_d	PRF_{dv}	PRF_{sdv}	IR	Baseline	letorr
NDCG@40	0.514	0.520	0.516	0.510	0.484	0.517
NDCG@100	0.489	0.497	0.491	0.489	0.440	0.491

initial ranking. This demonstrates that the proposed reranking features, though lightweight in computation, are still effective to be employed in the supervised learning-to-rerank method.

The neighborhood voting based features perform better than the reciprocal neighborhood voting based features. We argue that this is because we adopt the combination strategy for neighborhood construction, which makes the number of neighbors more adapted to the sample. In general, the reranking features which use the neighborhood structure perform better than the features which directly rely on the visual similarity. PRF_d achieves the worst performance among the 11 features. This further supports the hypothesis that features based on the neighborhood structure are less sensitive to the imperfections of visual similarity computation.

To further verify the effectiveness of the proposed features, in addition to training a reranking model using all 11 features, we also train reranking models by leaving each of the 11 features out. The result is shown in Table 2.5, from which we can make the following observations.

- Some of the reranking models trained by leaving one feature out are even better than the model using all of the features. For example, the reranking model trained without PRF_{dv} achieves 0.52 NDCG@40, which is higher than 0.517 using all the features. This suggests that there is redundancy among the 11 reranking features and that feature selection should be performed to preprocess the features in a general case for an improved performance.

- Some of the features, though with low individual performance, can complement the others so that their incorporation into the reranking model can still contribute to an improved performance. For example, the individual performance of PRF_d is only 0.355 in terms of NDCG@40, which is the worst among the 11 features. However, by incorporating it into the reranking model the performance is still improved by 0.5% compared with the model without it.

From the results discussed above we can say that the neighborhood voting based features perform better than the others, and that the selected 11 features are all useful, either directly or indirectly (in combination with other features).

2.5.5 Adapted Ranking SVM

As described in Section 2.3.3, we adapt the standard Ranking SVM algorithm by introducing an additional parameter α to model the importance of the initial ranking based feature. It is obvious that by setting α to 1 the standard Ranking SVM is achieved.

We vary α from 1 to 40 to see how the reranking performance is affected. The result is shown in Fig. 2.10. We can see that the reranking model with $\alpha = 1$, i.e., the standard Ranking SVM achieves the worst performance. This demonstrates that the adapted Ranking SVM is effective for learning the reranking model and indicates the range in which α should be selected.

2.6 Future Work

We see several possibilities to further explore and extend the learning-to-rerank paradigm, the major of which can be listed as follows.

- The existing unsupervised reranking methods can be employed to construct reranking features in the learning-to-rerank method. The advantage of the proposed 11 reranking features lies in the lightweight computational cost.

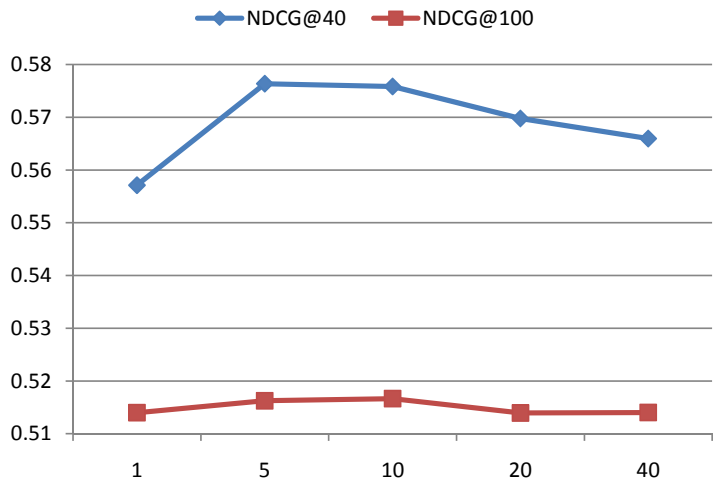

Figure 2.10: *The effects of alpha in the adapted Ranking SVM. When alpha =1 it becomes the standard Ranking SVM. The vertical axis is NDCG.*

However, in cases that the response time is not critical we can fuse multiple unsupervised reranking methods for a better ranking.

- The proposed learning-to-rerank method as well as the existing unsupervised reranking methods take only the relevance into consideration. However, the result diversity is also an important objective so that more informative search result can be provided to users. The learning-to-rerank framework makes it easy to take the result diversity into consideration by designing diversity-aware reranking features or a diversity-aware learning objective [134].

- As the experimental results reported in Table 2.5 suggest, the reranking feature selection can further improve the performance of the learned reranking model.

2.7 Conclusion

In this chapter, we introduced a supervised learning paradigm into visual search reranking to create a more robust reranking model. The idea leverages the advantages of both supervised concept-based search and unsupervised visual search reranking, while it does not suffer from scalability issues characteristic for concept-based search. To realize this idea, we proposed and formally defined a learning-to-rerank framework, which we implemented using the adapted Ranking SVM algorithm and 11 lightweight reranking features that encode the relevance between the textual query and visual documents. Experimental results obtained for

two representative image datasets indicate that the proposed supervised reranking paradigm can be considered a promising scheme for Web-scale image search.

Chapter 3

Prototype-based Image Search Reranking

[1]

In Chapter 2 we proposed a supervised reranking algorithm which shows improved effectiveness compared to the existing unsupervised methods. This approach, however, still leaves significant space for improvement, and this is mainly due to two simplifications we deployed in order to build and evaluate a proof of concept for supervised reranking. First, we assumed that all images in the top of the initial search results list are equally relevant to the query. Second, the reranking features were selected from a limited set of predefined features. The basic idea of the *prototype-based image search reranking* proposed in this chapter and generalizing the idea of supervised reranking is that we do not hypothesize about the relevance of the initial search result nor about the potentially useful reranking features. Instead, we learn both the relevance and the reranking features from the initial search result using a dedicated supervised learning framework.

[1] This chapter was published as: Linjun Yang, Alan Hanjalic, "Prototype-Based Image Search Reranking," IEEE Transactions on Multimedia 14(3-2): 871-882 (2012) [128].

3.1 Introduction

The existing web image search engines, including Bing [1], Google [2], and Yahoo! [3], retrieve and rank images mostly based on the textual information associated with the image in the hosting web pages, such as the title and the surrounding text. While text-based image ranking is often effective to search for relevant images, the precision of the search result is largely limited by the mismatch between the true relevance of an image and its relevance inferred from the associated textual descriptions [126].

To improve the precision of the text-based image search ranking, *visual reranking* has been proposed to refine the search result from the text-based image search engine by incorporating the information conveyed by the visual modality. Visual reranking has become a popular research topic in both multimedia retrieval and computer vision communities since it provides possibilities for considering the visual modality in the existing image search engines in a lightweight fashion and without incurring scalability issues. Moreover, apart from the image search scenario, visual reranking can also be used to improve the quality of the collected data in the process of automatically constructing training data from the web for object recognition [27][57].

While various techniques including clustering [37], topic modeling [29][27], SVM (Support Vector Machine) [121], graph learning [38][43][102], etc. have been investigated for the purpose of creating visual search rerankers, all of the existing reranking algorithms require a prior assumption regarding the relevance of the images in the initial, text-based search result. In the most widely used PRF (Pseudo Relevance Feedback) assumption [121][27][60][29][88][50], the top-N images of the initial result are regarded as pseudo relevant and used to learn a visual classifier for reranking. Even though the PRF-based reranking methods have been able to improve the precision over the initial text-based result in the past, the assumption that the top-N images are equally relevant can still be seen as too rigorous to be satisfied well by any arbitrary text-based image search engine. Since the text-based image search is far from perfect (which is the reason to perform the reranking in the first place), the top result will inevitably contain irrelevant images, which will introduce noise into the learning of reranking models and which may lead to sub-optimal search results being returned after reranking. In this sense, appropriately relaxing this assumption and redefining the reranking approach accordingly has the potential to further improve the precision of the visual reranking.

In this chapter we address this challenge by recalling the fact that image search engines usually optimize the system performance based on the relevance measures, such as NDCG (Normalized Discounted Cumulative Gain) [41], which tend to emphasize differently on the results at different ranks. Hence, it can naturally be assumed that the images in the top result of each query at different ranks have different probabilities to be relevant to the query. This should be incorporated into the reranking model for a more comprehensive utilization of the text-based search

result. Although this information has been investigated in previous work [102], the way in which it was utilized was rather ad hoc and therefore suboptimal. In this chapter, we propose a prototype-based method to learn a reranking function from human labeled samples, based on the assumption that the relevance probability of each image should be correlated to its rank position in the initial search result. Based on the images in the initial result, visual prototypes are generated that visually represent the query. Each of the prototypes is used to construct a meta reranker to produce a reranking score for any other image from the initial list. Finally, the scores from all meta rerankers are aggregated together using a linear reranking model to produce the final relevance score for an image and to define its position in the reranked results list.

The linear reranking model is learned in a supervised fashion to assign appropriate weights to different meta rerankers. Since the learned model weights are related to the initial text-based rank position of the corresponding image and not to the image itself, the reranking model is *query-independent* and can be generalized across queries. Consequently, the proposed reranking method can scale up to handle any arbitrary query and image collection, just like the existing visual reranking approaches, even though supervision is introduced.

The chapter is organized as follows. In Section 3.2 we briefly review the related work on visual reranking. In Section 3.3, we provide an overview of our proposed method and then focus in Section 3.4 on describing and discussing the key components of this method. The experimental results are presented and analyzed in Section 3.5, while Section 3.6 concludes the chapter with a brief overview of the main results of the chapter and the prospects for future work.

3.2 Related work

The methods for image search reranking can be classified into supervised and unsupervised ones, according to whether human labeled data has been used to derive the reranking model or not.

The unsupervised reranking methods do not rely on human labeling of relevant images but require prior assumptions on how to employ the information contained in the underlying text-based result for reranking. The most well-known assumption of this type is the PRF assumption. It considers the top ranked images in the text-based result as equally relevant to the query and uses them as positive samples for learning a reranking model [121][28][27][88][29]. While the reranking based on the PRF assumption has been demonstrated to often perform well, it suffers from a fundamental deficiency that we illustrate in Fig. 3.1. The diagram there shows the probability that an image at a given rank position in the initial, text-based search result is relevant to the query. We derived the probabilities from the search results we obtained on the Web Queries dataset (described in more detail in Section 3.5) for 353 representative image search queries. We can observe that on this dataset only 35 top-ranked images with the relevance proba-

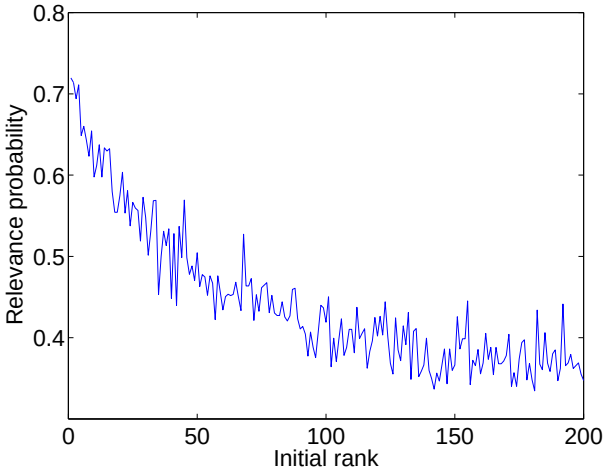

Figure 3.1: *The relevance probability of images at different rank positions of the text-based search result, derived from the statistics in the Web Queries dataset comprising 353 representative image queries. The relevance probability is computed as the average relevance of the images ranked at this position based on the ground-truth.*

bility above 0.5 could be considered relevant, though noisy, and used for learning the reranking model. This number of relevant images is, however, too small to learn a robust model. Using more images is the only alternative, but still not a good one. Lower-ranked images may namely be irrelevant and therefore introduce more noise in the model learning process.

The other widely-adopted image search reranking assumption is the cluster assumption, which says that the visually similar images should be ranked nearby [102]. Based on this assumption, various graph-based methods [38][43][102][103] have been proposed to formulate the image search reranking problem. The main deficiency of this assumption is that it makes the visual similarity of images equal to the similarity of their relevance to the query. In addition, it omits to identify two images as equally relevant to the query if they are insufficiently visually similar to each other. Although effort has been invested in the selection of visual features and similarity criteria that map visual similarity into relevance [113], this semantic gap has not yet been successfully bridged.

A straightforward way of coping with the deficiencies of unsupervised reranking methods described above is to rely on manual relevance labeling of a training data set, that is, to introduce human supervision in the reranking process. Such supervision, however, needs to be embedded in such a way that the learned reranking model can scale up beyond the training data collection and queries used in the learning step. Hence, relevance feedback based approaches [142][101] cannot be applied since there query-specific models will be learned, which requires

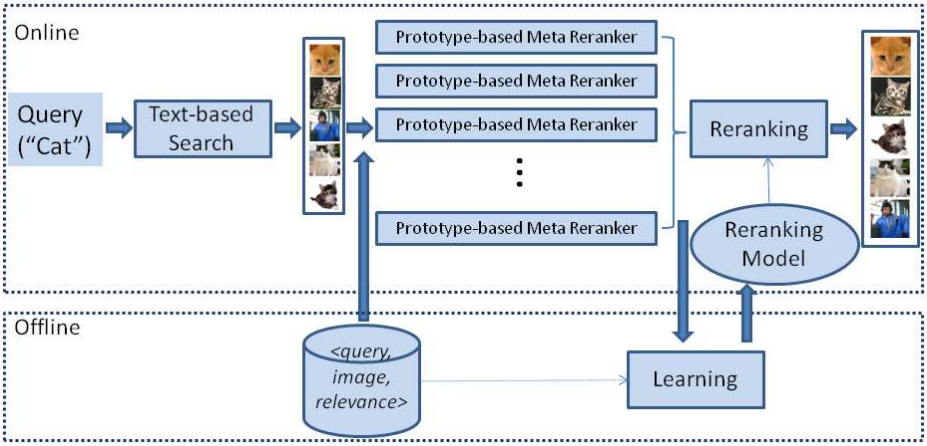

Figure 3.2: *Overview of the proposed prototype-based visual reranking framework.*

labeling from users for each submitted query. In view of this, the challenge of supervised reranking is to design query-independent reranking models based on query-dependent reranking features. These features typically model the pairs of a textual query and an image document taken from the initial, text-based search result. Recent successful attempts in this direction have been made by Yang and Hanjalic [126] and Krapac et al. [50]. While human supervision helps alleviate the problems of unsupervised methods, the existing methods are still far from optimal. The reranking features used in [126] and [50] are still designed based on the PRF and cluster assumptions. In addition, although in [126] the contribution of images into the reranking features varies with their initial rank positions, this variation is based on hand-crafted rules. These rules may work well for some data collections and text-based search engines, but their suitability is difficult to be shown in a general case. The method proposed in this chapter makes a further step in the development of supervised, but scalable visual reranking systems by explicitly targeting the improvement in the robustness and reliability of the learned reranking model.

3.3 Prototype-based reranking

3.3.1 System framework

As illustrated in Fig. 3.2, the proposed prototype-based reranking method consists of an online and an offline step.

In the online part, when a textual query is submitted to the image search engine by a user, initial search is performed using any contemporary text-based search technique [67]. Then, L visual prototypes are generated and for each

prototype a meta reranker is constructed. The construction of meta rerankers is explained in detail in Section 3.4. Then, for each of the top-N images in the initial search result, an L-dimensional score vector is obtained comprising the scores from all L meta rerankers when applied to that image. Finally, the score vector is used as input to a reranking model, which has already been trained offline, to estimate the ranking scores in the reranked image search list.

The offline component is devoted to learning the reranking model from user-labeled training data. Since the learned model will be used for reranking the text-based search results, the training set is constructed from these results through the following steps. First, several representative queries sampled from the query log are selected. Then, using these queries the top N images are retrieved from the text-based image search engine and downloaded for processing. Finally, for each query-image pair, people are invited to label the relevance between them to form the ground-truth. After the training data is collected, we can compute the score vector from the meta rerankers, as mentioned in the online part, for each image and the corresponding query. Then the reranking model is learned and stored in the memory to be used in the online part for responding to users' submitted queries.

3.3.2 Learning a reranking model

The linear reranking model adopted in this chapter is learned by estimating the weights of the combined scores coming from different meta rerankers. This problem can be addressed using a learning-to-rank method [59], by regarding the score vector as the ranking feature of an image.

Ranking SVM [44] is among the most popular learning-to-rank algorithms and we also adopt it in this chapter. This algorithm adapts the widely-used SVM classifier to handle the ranking problem. The basic idea is to decompose a ranking into a set of pair-wise preferences and then to reduce the ranking-learning problem into a pair-wise classification problem. Ranking SVM learns the ranking model by solving an optimization problem that can be defined as follows:

$$\begin{aligned} \min\ & \frac{1}{2}W^T W + C\sum \xi^i_{jk} \\ s.t.\ & \forall q_i, I_j \succ I_k : W^T(M(I_j) - M(I_k)) \geq 1 - \xi^i_{jk} \\ & \forall i, j, k : \xi^i_{jk} \geq 0, \end{aligned} \tag{3.1}$$

where W is the model weight vector, C is the parameter to trade-off the loss and the regularization, $M(I_j)$ is the score vector from the L meta rerankers for the image I_j, ξ^i_{jk} is the slack variable, and $I_j \succ I_k$ means that image I_j is more relevant than I_k for the query q_i.

Standard efficient approaches to learning an SVM classifier, such as SMO (Sequential Minimal Optimization) [78], can be directly employed for learning

the Ranking SVM. Moreover, a fast algorithm, e.g., the cutting-plane algorithm [45], can be adopted to speed up the training of a linear Ranking SVM.

3.3.3 Discussion

The reason why the learned reranking model described above can be generalized across queries beyond those used for the training is that the model weights are not related to specific images but to their rank positions in the text-based search result. The separation of the model weights from specific images is the key to ensure that the reranking model only needs to be learned once and can then be applied to any arbitrary query.

The existing learning-to-rerank methods, including the supervised-reranking [126] and query-relative classifier [50], design the reranking model based on the hand-designed ranking features defined at a higher abstraction level or on the ordered visual words, respectively. Compared to them, the prototype-based learning to rerank method learns how likely the images at each of the ranked position in the text-based result are to be relevant to the query. In other words, the method directly learns the characteristics of the underlying text-based image search engine and requires less expert input in terms of the reranking feature design and a more relaxed assumption on the underlying text-based search than, for instance, [126] and [50]. Consequently, the prototype-based reranking method can be expected to generalize even better over a broad set of queries and perform well for any underlying text-based search engine.

3.4 Constructing meta rerankers

One of the key steps in the prototype-based image search reranking method is the construction of meta rerankers. Given a prototype P_i and a set of N images $\{I_j\}_{j=1}^{N}$, the task here is to compute the ranking scores $\{M(I_j, P_i)\}_{j=1}^{N}$ for these images based on the prototypes. The computed scores are then used as input for the reranking model to estimate the ultimate ranking scores to determine the rank position of the images in the reranked result. In the following, we propose three types of meta rerankers, depending on how the prototypes are generated from the initial text-based search result.

3.4.1 Single-image prototype

A straightforward way to generate a set of prototypes is to select top L images from the text-based result, as illustrated in Fig. 3.3. If we denote this set as $\{P_i^S\}_{i=1}^{L}$, then the meta reranker can be built simply based on the visual similarity $Sim(I_j, P_i^S)$ between the prototype P_i^S and the image I_j to be reranked:

$$M^S(I_j, P_i^S) = Sim(I_j, P_i^S). \tag{3.2}$$

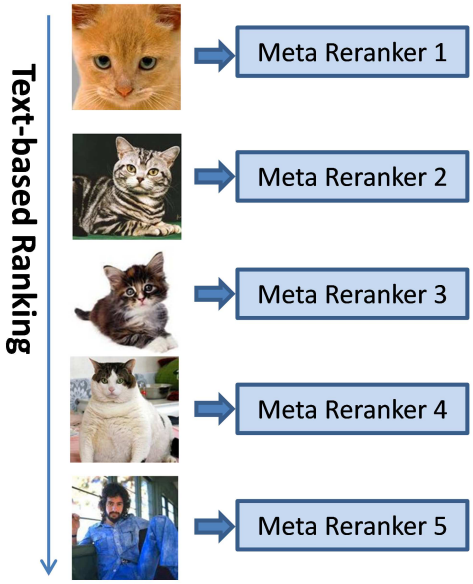

Figure 3.3: *Illustration of constructing meta rerankers using single-image prototype.*

The score vector aggregating the values (3.2) from all L meta rerankers is then used as input to the linear reranking model in order to compute the definitive ranking score for image I_j :

$$R^S(I_j) = \sum_{i=1}^{L} w_i \times Sim(I_j, P_i^S), \tag{3.3}$$

where w_i are the individual weights from the model weight vector W.

RRC assumption

If we recall the discussion of the previous work in Section 3.2, and in particular in relation to the PRF assumption serving as the basis of many existing visual reranking methods, it can be said that the proposed reranking method using single-image prototypes is also based on an assumption, namely that the relevance of an image should be correlated to its rank position in the text-based result. We refer to this assumption as the *RRC (Relevance-Rank Correlation) assumption*, which can also be seen as a relaxed version of the PRF assumption. Compared to the rerankers based on the PRF assumption, the RRC-based reranking methods are expected to be more robust to imperfection and unreliability of the text-based search result, since the relevance-rank correlation is actually reflected in the objective of a search engine.

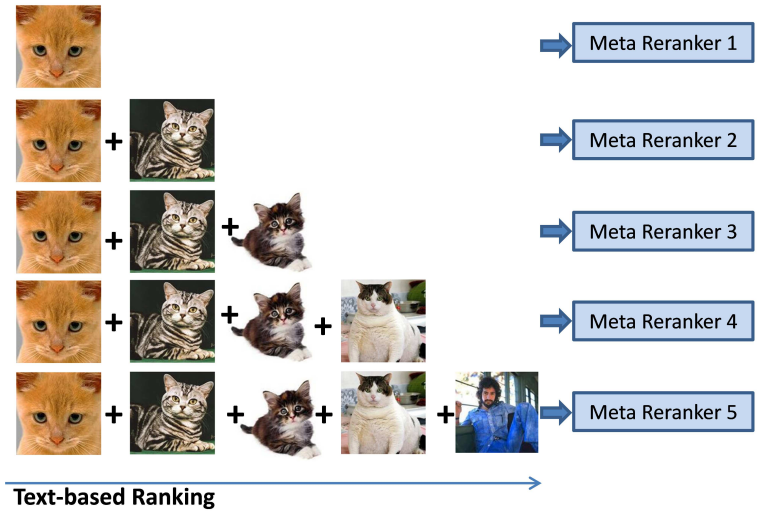

Figure 3.4: *Illustration of constructing meta rerankers using multiple-average prototype.*

Query independence

It is interesting to observe that the model (3.3) is similar in form to the kernel model trained by PRF-based methods using the top images in the text-based result. However, while PRF-based methods learn query-specific models separately for different queries, our approach learns the models in a query-independent way. As discussed before, the advantage of the query-independent learning is that it can leverage the labeled data from a limited number of queries to train a unified model which can then be generalized across a broad range of queries. In this way, introducing supervision in the learning process does not jeopardize scalability.

3.4.2 Multiple-average prototype

While the RRC assumption introduced in Section 3.4.1 is more powerful than the PRF assumption, the noisiness of the relevance distribution in Fig. 3.1 indicates that the relevance-rank correlation criterion is not necessarily fulfilled at all rank positions in the initial text-based search result. For example, in Fig. 3.1, the relevance probability at rank 41 is only 0.448 while that at rank 44 is 0.537. In order to leverage the effect of possible correlation distortions at individual rank positions, instead of considering a single image as a prototype, one could also consider a "bag" of images taken from the neighboring rank positions.

Following this rationale, as an alternative to the prototype definition in Section 3.4.1, we now construct a prototype P_i^{MA} by first selecting the top L images in the initial search result list and then by cumulatively averaging the features of

all images ranked starting from the the topmost position to the position i, as illustrated in Fig. 3.4. In other words, the prototype P_i^{MA} can be defined as

$$P_i^{MA} = \frac{1}{i}\sum_{j=1}^{i} I_j. \tag{3.4}$$

Here the summation indicates the process on suitable features of I_j.

Then, the prototypes (3.4) can be employed to compute the scores of individual meta rerankers by again computing the visual similarity between a prototype and the image to be reranked:

$$M^{MA}(I_j, P_i^{MA}) = Sim(I_j, P_i^{MA}). \tag{3.5}$$

Bag-wise RRC assumption

In relation to the RRC-based reranking introduced in the previous section, we can say that the meta rerankers (3.5) are generated under the *bag-wise RRC (BRRC) assumption*, which states that a rank position should be correlated with a bag of images, instead of an individual image. This is equivalent to smoothing out the noise from the probability distribution in Fig. 3.1 through which more robust reranking models can be achieved.

Fig. 3.5 shows the relevance probability of each bag at different ranks, which is estimated as the mean relevance of all images contained in that bag. By comparing Fig. 3.5 and Fig. 3.1 it can be observed that the relevance of a bag is better correlated with the rank than the relevance of an individual image.

The BRRC-based reranking approach proposed in this section is a straightforward way to utilize the BRRC assumption by using the average image features as the representation of the bag. In Section 3.4.3 we will present another approach based on the BRRC assumption, where a visual classifier from each bag is trained as the representation of that bag.

Analysis

In this section, we will analyze the properties of the reranking method based on the multiple-average prototype with a special case of similarity measure based on the dot product. While we notice that the following mathematical derivation may not be applicable for other categories of similarity functions, we believe the properties of multiple-average prototype obtained from this analysis are generally useful. By using the dot product similarity the corresponding meta reranker can be written as

$$M^{MA}(I_j, P_i^{MA}) = \frac{1}{i}\sum_{k=1}^{i} Sim(I_k, I_j). \tag{3.6}$$

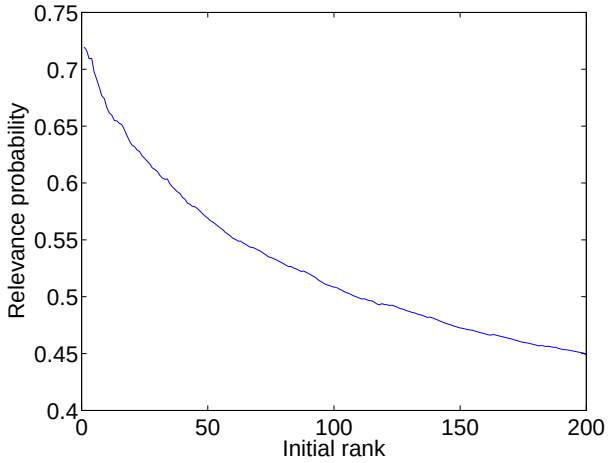

Figure 3.5: *The relevance probability of bags of images at different rank positions of the text-based search result, derived from the statistics in the Web Queries dataset comprising 353 representative image queries. The relevance probability is computed as the average relevance of the images ranked at and before this position based on the ground-truth.*

Integration of (3.6) into the linear reranking model leads to the following expression:

$$\begin{aligned} R^{MA}(I_j) &= \sum_{i=1}^{L}(w_i \times \frac{1}{i}\sum_{k=1}^{i} Sim(I_k, I_j)) \\ &= \sum_{i=1}^{L} \alpha_i \times Sim(I_i, I_j), \end{aligned} \tag{3.7}$$

where

$$\alpha_i = \sum_{k=i}^{L} \frac{w_k}{k}. \tag{3.8}$$

The above expressions transform the model based on a multiple-average prototype onto the model based on a single-image prototype, however, with different weights.

The reranking model based on a multiple-average prototype has three important properties. The first is that the weights of images ranked higher in the text-based search result will be larger than that of the images ranked lower:

$$\alpha_i >= \alpha_j \text{ for } i < j. \tag{3.9}$$

This property can easily be derived from Eqn. (3.8). It states that the ranking in the text-based search result represents the ordering of the importance for each individual image to be used as a prototype for reranking. In other words, the

reranking based on a multiple-average prototype will rely more on the initial text-based result than that based on a single-image prototype.

To derive the second and the third property, we write the model weights W as

$$w_i = i \times \sum_{k=i}^{L} (-1)^{k-i} \alpha_i. \tag{3.10}$$

Then we integrate this formula into the formulation of Ranking SVM as defined in Eqn. (3.1) and obtain

$$\begin{aligned} \min \ & \frac{1}{2} \sum_i |i \times \sum_{k=i}^{L} (-1)^{k-i} \alpha_i|^2 + C \sum \xi_{jk}^i \\ s.t. \ & \forall q_i, I_j \succ I_k : A^T (M(I_j) - M(I_k)) \geq 1 - \xi_{jk}^i \\ & \forall i, j, k : \xi_{jk}^i \geq 0, \end{aligned} \tag{3.11}$$

where A is the vector of α_i.

From the above expression we can see that the regularization of each model parameter α_i is weighted by its rank. Hence, the second property of the reranking based on a multiple-average prototype is that the different α parameters have different flexibility to find the optimal value. The parameters corresponding to higher ranks (smaller i) have a larger solution space, and vice versa. This has a similar effect as the feature balancing strategy in the supervised-reranking method to emphasize the important features a priori [126]. For the reranking method based on a multiple-average prototype, the higher the image is in the text-based ranking the more important it is for reranking.

The third property is also derived from the regularization. The reranking model in Eqn. (3.11) not only regularizes the solution space of model parameters α, but also regularizes to make the images at adjacent ranks have similar weights. Combining it with the first and second property, we can conclude that the learned weights for individual images by the reranking based on a multiple-average prototype will decline gradually with the decreasing ranks. This may make this reranking model less aggressive and more robust than the one based on a single-image prototype. Meanwhile, it makes the reranking model learned for the multiple-average prototype hardly over-fitting to the training queries.

3.4.3 Multiple-set prototype

The multiple-set prototype P_i^{MS} at rank i is defined as a bag of images ranked from the topmost position to the rank i, as illustrated in Fig. 3.6.

$$P_i^{MS} = \{I_j\}_{j=1}^{i}. \tag{3.12}$$

The multiple-average prototype presented in Section 3.4.2 is the average of features for the images in the multiple-set prototype and can be seen as a special

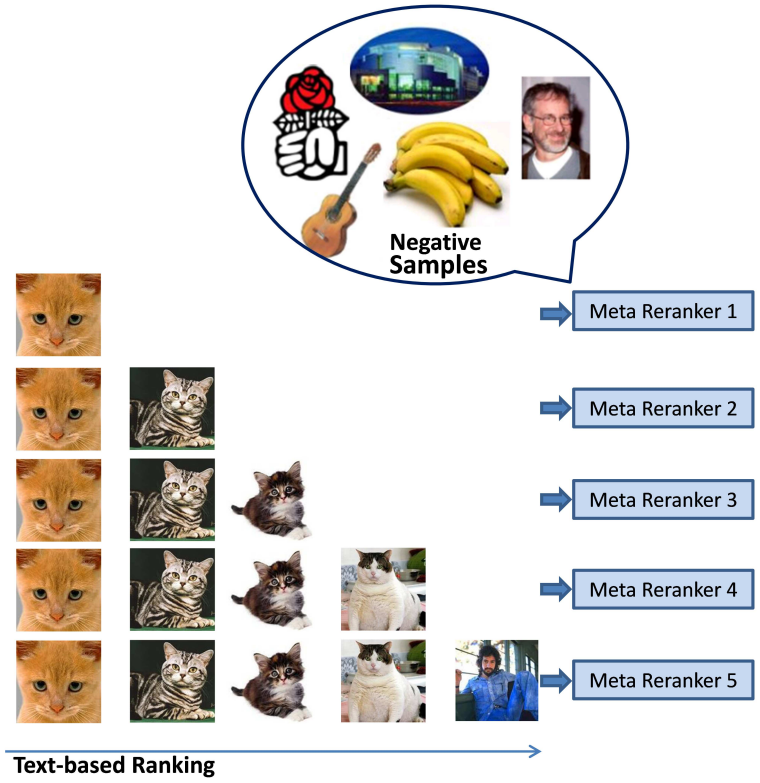

Figure 3.6: *Illustration of constructing meta rerankers using multiple-set prototype.*

case of this prototype. The multiple-set prototype is a more flexible representation satisfying the bag-wise RRC assumption, which can support the development of other types of meta rerankers.

Given a multiple-set prototype P_i^{MS}, we can learn a visual classifier by regarding all the images in P_i^{MS} as positive samples, which is then employed as meta reranker and the prediction score is used as the meta reranking score.

Since a discriminative learning method is usually more effective for learning a visual model, we adopt SVM [99] in this chapter. However, it needs not only positive samples but also negative samples. We propose the following two strategies to select negative samples.

- **Background images.** The advantage of selecting the background images as negative samples is that they are very unlikely to be relevant to any query of interest. In this chapter, we select the images which are ranked in the bottom for each query as the background.

- **Random images.** The other strategy of selecting negative samples is to

use the randomly sampled images from the entire database. The advantage of selecting random images as negative samples is that we can construct multiple sets of negative samples, so as to de-correlate different meta rerankers.

The meta reranker with a multiple-set prototype can be defined as follows:

$$M^{MS}(I_j, P_i^{MS}) = p(I_j|\hat{\theta}), \tag{3.13}$$

where $\hat{\theta}$ is the learned model and

$$\hat{\theta} = \arg\max_{\theta} p(P_i^{MS}|\theta). \tag{3.14}$$

3.5 Experiments

In this section, we demonstrate the effectiveness of the proposed prototype-based image search reranking method by means of an experimental study performed on the publicly available Web Queries dataset. We refer to our three reranking methods: single-image prototype, multiple-average prototype, and multiple-set prototype, proposed in Section 3.4, as *Prototype-Single*, *Prototype-Average*, and *Prototype-Set*, respectively.

3.5.1 Experimental setup

To make the experimental results reported in this chapter reproducible, we used the publicly available Web Queries dataset[2] comprising a large amount of representative and diverse image search queries. The dataset contains in total 353 queries. For each query, tens or hundreds of images are retrieved and downloaded using a web search engine, which resulted in a total of 71478 images. For each query and an image in its text-based search result, a binary relevance is labeled as the ground-truth.

To illustrate the effectiveness of the proposed method, we compare it with the text-based search baseline from the search engine as well as the state-of-the-art supervised and unsupervised reranking methods. The supervised approaches include the recently proposed supervised-reranking[3] [126] and query-relative classifier [50]. The unsupervised reranking methods include the PRF reranking [121], random walk reranking [38][43], and Bayesian reranking [102][103]. In the PRF reranking, top-ranked images in the initial search result for a given query are used as positive samples and the negative samples are selected from the background images. For Bayesian reranking, the best performing local learning consistency and pair-wise ranking distance are used. As suggested in [50], 400 ordered visual words are used to construct the binary features for query-relative classifier.

[2] http://lear.inrialpes.fr/∼krapac/webqueries/webqueries.html

[3] To differentiate the specific method called "supervised reranking" in [126] from the general meaning of supervised reranking, we will use "supervised-reranking" to denote the method in [126].

The reranking models in the supervised reranking methods including those from [126][50] and the one proposed in this chapter are trained using Ranking SVM [44]. To validate the parameter C in Ranking SVM, the entire dataset is randomly split into 10 folds to generate the training, validation, and test data. The reranking methods are tested in a round robin way for 10 times. At each time, one among the 10 folds is used for testing, 8 folds for training, and the rest one for validating the parameter C.

For the purpose of computing visual similarity between images, the SIFT features [61] with dense sampling are extracted from the images and then quantized together with the spatial layout to represent the images as bags of visual words [110]. For the methods including random walk reranking, Bayesian reranking, supervised-reranking, and our proposed method, histogram intersection is used as the similarity measure between two images. For PRF reranking and the variant of our method based on the multiple-set prototype, the linear kernel is adopted for SVM due to efficiency reasons. In addition to the visual features, 154-dimensional textual ranking features are extracted for each query-image pair from the text associated with the images, according to a common text search approach[4]. The textual ranking features will be combined with the visual ranking features to achieve a better performance. The reranking using only the 154-dimensional textual ranking features is also reported.

The SVMLight software [99] is employed to learn the classifiers for PRF reranking and for constructing the meta rerankers for the reranking with multiple-set prototypes. Since cross-validation is time-consuming, the default value of the parameter C estimated by the software is adopted.

All the images in the text-based result for a query in the Web Queries dataset are involved in the reranking process. For *Prototype-Single* and *Prototype-Average* the number of prototypes is set as the number of images in the text-based result for a query, while for *Prototype-Set* we use at most top 100 images to build 100 meta rerankers. When constructing the meta rerankers for *Prototype-Set* We use 352 samples as negative samples that are drawn from a different query than the one being used.

Both Average Precision (AP) [8] and Normalized Discounted Cumulative Gain (NDCG) [41] are adopted to measure the ranking performance. AP is defined as the average of precisions at various recall levels. The APs for all queries in the dataset are averaged to obtain the Mean Average Precision (MAP). NDCG emphasizes more on the relevance of top results through discounting the gain by the ranked position, which is defined as

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \tag{3.15}$$

[4]http://research.microsoft.com/en-us/projects/mslr/feature.aspx

Table 3.1: *Performance comparison of various reranking methods. The numbers in the brackets are the relative improvements of various methods over the Text-baseline.*

Methods	MAP	NDCG@10	NDCG@40
Text-baseline	0.569	0.682	0.633
Text-ranking	0.594 (+4.39%)	0.684 (+0.29%)	0.649 (+2.53%)
PRF [121]	0.658 (+15.64%)	0.772 (+13.20%)	0.718 (+13.43%)
Random walk [38][43]	0.641 (+12.65%)	0.766 (+12.32%)	0.704 (+11.22%)
Bayesian [102]	0.643 (+13.01%)	0.766 (+12.32%)	0.709 (+12.01%)
Supervised-reranking [126]	0.665 (+16.87%)	0.769 (+12.76%)	0.733 (+15.80%)
Query-relative [50]	0.666 (+17.05%)	0.768 (+12.61%)	0.729 (+15.17%)
Prototype-Single	0.678 (+19.16%)	0.804 (+17.89%)	0.750 (+18.48%)
Prototype-Average	0.669 (+17.57%)	0.794 (+16.42%)	0.742 (+17.22%)
Prototype-Set	0.703 (+23.60%)	0.826 (+21.11%)	0.777 (+22.75%)
Prototype-All+Query-relative	0.706 (+24.08%)	0.823 (+20.67%)	0.779 (+23.06%)
Prototype-Set+Text	0.714 (+25.48%)	0.835 (+22.43%)	0.787 (+24.33%)

where

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2(i+1)}, \tag{3.16}$$

r_i is the human judged relevance of the corresponding image and IDCG@k is a normalization term used to scale the NDCG between 0 and 1.

3.5.2 Performance comparison

The proposed three variants of the prototype-based reranking method, including the one based on a single-image prototype (*Prototype-Single*), multiple-average prototype (*Prototype-Average*) and a multiple-set prototype (*Prototype-Set*) are compared with the baseline from the text-based search engine (*Text-baseline*), textual ranking based on a learned ranking model with the 154-dimensional textual ranking features (*Text-ranking*), and the state-of-the-art unsupervised and supervised visual reranking methods, including supervised reranking (*Supervised-reranking*) [126], query-relative classifier (*Query-relative*) [50], PRF reranking (*PRF*) [121], random walk reranking (*Random walk*) [38][43], and Bayesian reranking (*Bayesian*) [102][103].

Table 3.1 shows the performance comparison of the above-mentioned reranking methods, in terms of MAP, NDCG@10, and NDCG@40. It can be clearly observed that all the visual reranking methods outperform the *Text-baseline* with performance improvements larger than 12%, and *Text-ranking* with improvements above 8%, in terms of MAP. This demonstrates the effectiveness of the visual reranking concept. Furthermore, all supervised reranking methods outperform the unsupervised ones, which once again justifies the integration of a manual relevance labeling step in the reranking process.

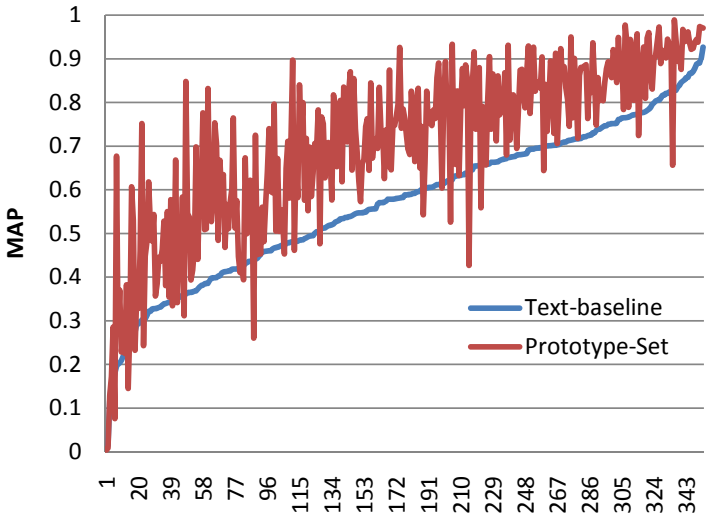

Figure 3.7: *Performance comparison of* Prototype-Set *and* Text-baseline *from the search engine. The queries are arranged in the ascending order of the performance of* Text-baseline. *The x axis are the queries labeled by the corresponding ranks.*

Among the three kinds of supervised reranking methods, all three variants of the proposed prototype-based reranking method outperform *Supervised-reranking* and *Query-relative* methods. Since the learning approach and the visual feature representation are nearly the same for these reranking methods, this result shows that the prototype-based meta rerankers are most successful in effectively utilizing the information extracted from the text-based search result. This also demonstrates the effectiveness of the proposed RRC and bag-wise RRC assumptions. Since the *Query-relative* classifier uses the average of the visual features from top images as the visual representation of the query, the performance may be influenced by the outliers in the top result and the differentiation of images at different ranked positions is not taken into consideration. On the contrary, the prototype-based reranking approach is hardly affected by outliers, since the weights for different images are learned from human-labeled data. Compared to the *Supervised-reranking* approach, which extracts eleven carefully designed reranking features based on the domain knowledge from the image search context, the prototype-based reranking method can be thought of as to be more generalized since the knowledge about the image search engine is discovered automatically.

Among the three variants of the prototype-based reranking method, *Prototype-Set* achieves the best performance, which improves 23.6% in terms of MAP over the *Text-baseline*. The reason may lie in the fact that *Prototype-Set* more comprehensively utilizes the information in the text-based result through learning

visual classifiers. A comparison of the MAP performance between *Prototype-Set* and *Text-baseline* in Fig. 3.7 shows that search performance is improved for 93% (327/353) of the queries by the *Prototype-Set* reranking.

For the evaluation of the computational cost of the proposed approaches, we mainly focus in the chapter on the overall online computational cost that directly affects the query response time. The computational cost of reranking images for a query is 12.47s using *Prototype-Set* on a single CPU in our workstation, which is mainly due to the training of meta-rerankers. While this is acceptable for the tasks including collecting training data from web, it is not appropriate for real-time web image search tasks. However, the speed can be traded off with the precision. The reranking time can be reduced to 3.6s by using less negative samples (100 negative samples), and can be further reduced to 0.78s by constructing meta rerankers at every five images. The corresponding MAP values are 0.687 and 0.684, respectively, which are still better than those of other methods.

The combination of the meta rerankers constructed by the three variants of the prototype-based reranking method and the features in query-relative classifiers (*Prototype-All+Query-relative*) does not perform better than only using *Prototype-Set*. This suggests that *Prototype-Set* already exploits all the visual information which can be mined from the text-based search result and that the information discovered by *Prototype-Single*, *Prototype-Multiple*, and even *Query-relative* does not reveal any new aspects of influence for reranking.

To complete the experimental study, we also build a new reranking method, *Prototype-Set+Text*, by integrating the meta rerankers of the best performing *Prototype-Set* and the 154-dimensional textual ranking features. This hybrid method outperforms all the others. Its ranking performance arrives at 0.714 in terms of MAP and achieves 25.48% improvement over the *Text-baseline*. Moreover, *Prototype-Set+Text* improves the results for 95% (334/353) of the queries over the *Text-baseline*.

Sample results from *Prototype-Set+Text* and the *Text-baseline* are illustrated in Fig. 3.8. We can see that the proposed method can indeed learn meaningful visual models from the text-based search result in a query-independent way to boost the ranking performance. For most of the queries, e.g., "forbidden city" and "white house", the ranking performance is greatly improved. Even when the text-based ranking is bad, the proposed method can still discover useful information for reranking. For example, for the query "comics page", the MAP for *Text-baseline* is only 0.161. However, the reranking can still boost the performance by improvement of 63% since the relevant images exhibit common patterns while noisy images are scattered. In few cases (less than 5%) where the text-based result is poor and the noisy images are visually similar to the relevant ones, the reranking will degrade the performance. For example, for the query "will smith", even though it can be discovered that the query is about people, the visual features cannot distinguish persons from each other.

Figure 3.8: *Sample results of (a)* Text-baseline *and (b)* Prototype-Set+Text. *The numbers 1 or 0 below the images are the manually labeled relevance values serving as ground truth.*

3.5.3 Analysis

We can observe from Table 3.1 that *Prototype-Average* performs slightly worse than *Prototype-Single*. This seems to be in contradiction with the intuition since *Prototype-Average* is based on a more robust assumption. However, the analysis in Section 3.4.2 indicates that this result is reasonable since *Prototype-Average* is more moderate in changing the text-based search result than *Prototype-Single*. Fig. 3.9 shows a per-query performance comparison between *Prototype-Single* and *Prototype-Average* with the text-based search result serving as baseline. This comparison shows that while *Prototype-Average* is less effective in boosting the reranking performance, it is also more robust in the cases when reranking may lead to performance degradation, like in the last examples discussed in the previous section.

Figure 3.10 illustrates the respective performance of the meta rerankers constructed by *Prototype-Set* using bags at different ranks. Two observations can be

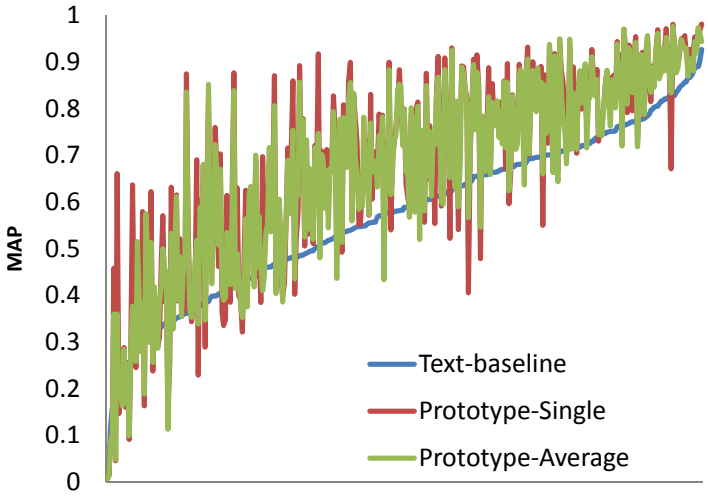

Figure 3.9: *MAP comparison of* Prototype-Single *and* Prototype-Average *methods.*

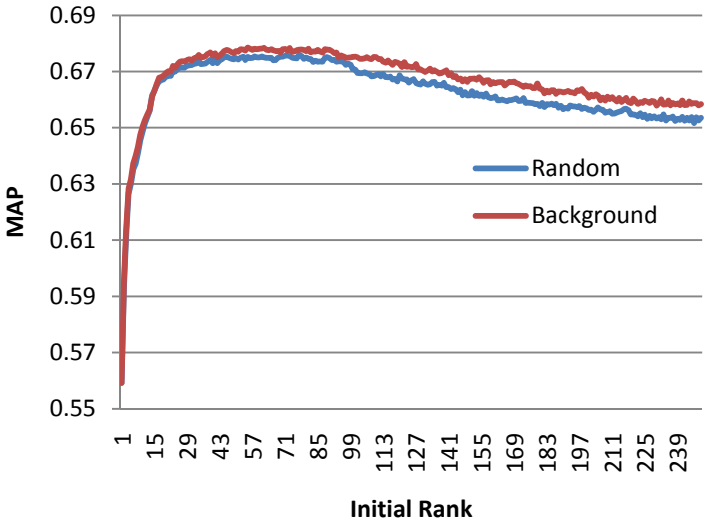

Figure 3.10: *Performance of various Multiple-set prototype based meta rerankers corresponding to different ranked positions in the text-based search result.*

made. First, the data size is the most important factor influencing the performance of meta rerankers when the data size is small. For the meta rerankers with the background images strategy, the MAP quickly improves from 0.559 to 0.632 when the number of positive samples increases from 1 to 5. It arrives at a peak at 0.679 when top 55 images are used to construct the meta reranker. After that

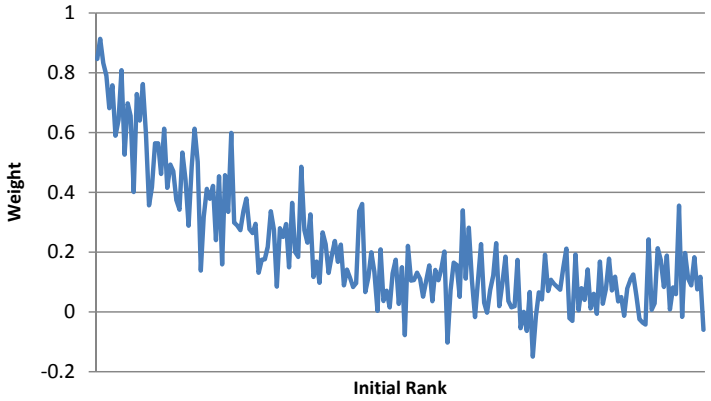

Figure 3.11: *Visualization of the learned model weights w for* Prototype-Single *method.*

the performance starts to decline with more images being incorporated into the learning process and arrives at 0.658 when all top images are utilized. This leads to the second observation that the data quality is the most important factor when the data size is sufficient. Since finding a trade-off between the data quality and data size is a challenging problem, *Prototype-Set* addresses this problem by learning weights to combine the meta rerankers with high data quality or large data size, to more comprehensively utilize the information contained in the text-based search result.

The comparison of the meta rerankers using random images and background images shown in Fig. 3.10 demonstrates that the two strategies are comparable, although background images perform slightly better than random images. The MAP of *Prototype-Set* with background images (0.705) is also slightly better than that with random images (0.702).

The learned reranking models are visualized in Fig. 3.11, Fig. 3.12, and Fig. 3.13 for *Prototype-Single*, *Prototype-Average*, and *Prototype-Set*, respectively. We can see that the model weights in *Prototype-Single* tend to decrease with the decline of the ranked positions in the text-based search result. While the decrease is not strict, it is basically accordant with the relevance probability as shown in Fig. 3.1. Figure 3.12 shows that *Prototype-Average* exhibits a similar trend, that is, the weights for top-ranked multiple-average prototypes in the text-based result tend to be larger than those ranked lower. As the weights for the multiple-average prototypes in *Prototype-Average* can be transformed to be the α weights on individual images using Eqn. (3.8), the α values are computed and shown in Fig. 3.14. The α values decrease strictly and smoothly with the ranked position and become nearly zero after rank 78, which demonstrates the correctness of the derived three properties of *Prototype-Average* in Section 3.4.2. From this result and the former analysis we can hypothesize that although *Prototype-Average* cannot perform better than *Prototype-Single* in terms of average performance, it is more

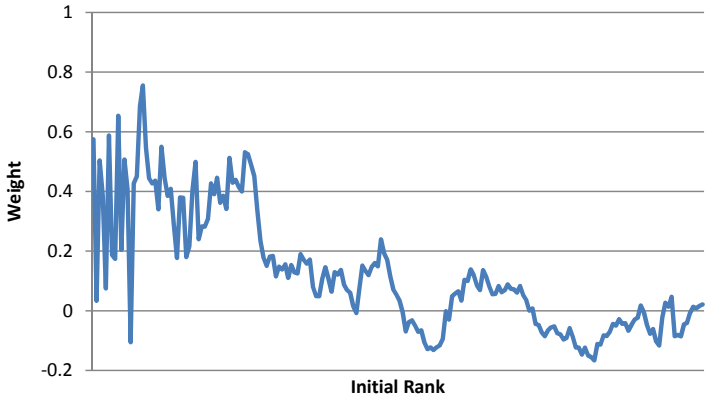

Figure 3.12: *Visualization of the learned model weights w for* Prototype-Average *method.*

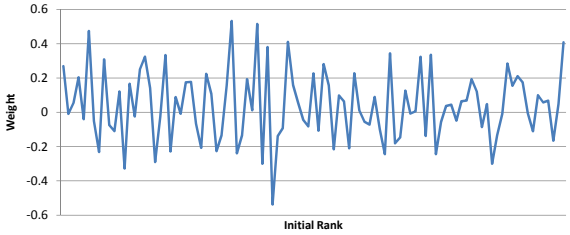

Figure 3.13: *Visualization of the learned model weights w for the* Prototype-Set *method.*

robust to noise and hardly suffers from the over-fitting problem. The learned weights for *Prototype-Set* as shown in Fig. 3.13 have a more complex relationship with the rank positions since the meta rerankers learned from the bags at different ranks are highly correlated. A conclusion that can be drawn from this is that the learning process is rather important since we cannot simply design a function to estimate the model weights from the rank positions.

We further studied the effect of two parameters in *Prototype-Set*, i.e., the number of meta rerankers L and the number N of top images to be reranked, on the reranking performance. We can see in Fig. 15 that the variation in the performance reduces for $L >= 30$ and that the performance exceeds 0.7 for $L >= 100$. Since increasing L leads to larger computational costs, we select $L = 100$ in our experiments as a good representative value for our experiments.

Intuitively, involving more top images from the initial result into the reranking process should lead to an increase in the reranking performance. Fig. 3.16 shows the performance of *Prototype-Set* when applied to different numbers of top images, from which we can see that the MAP keeps improving when N increases. This

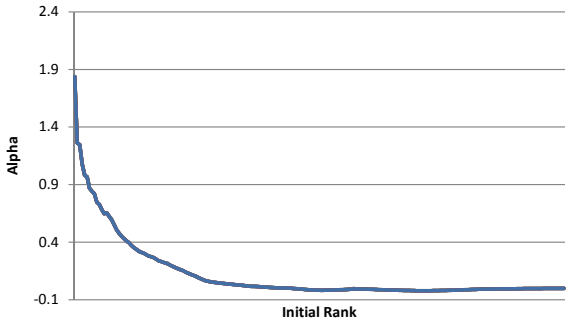

Figure 3.14: *Visualization of the parameters α for the* Prototype-Average *method.*

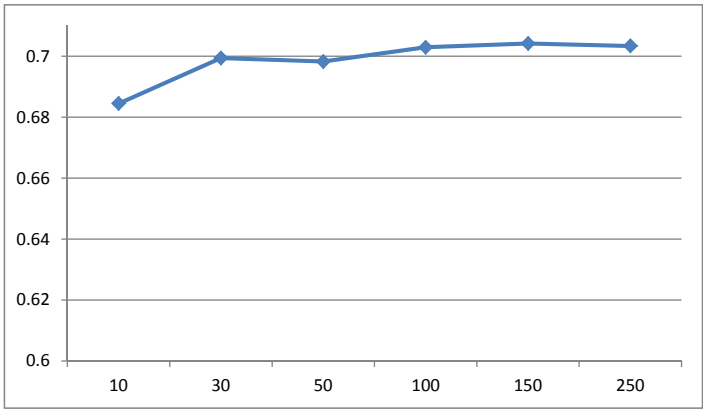

Figure 3.15: *The MAP of* Prototype-Set *for different numbers L of meta rerankers.*

basically demonstrates the correctness of the intuition. On the other side, we may also intuitively hypothesize that more images to be reranked could lead to increased unreliability of the reranked result. However, we did not observe this in our results.

3.6 Conclusions

In this chapter, we proposed a prototype-based reranking framework, which constructs meta rerankers corresponding to visual prototypes representing the textual query and learns the weights of a linear reranking model to combine the results of individual meta rerankers and produce the reranking score of a given image taken from the initial text-based search result. The induced reranking model is learned in a query-independent way requiring only a limited labeling effort and being able to scale up to a broad range of queries. The experimental results on

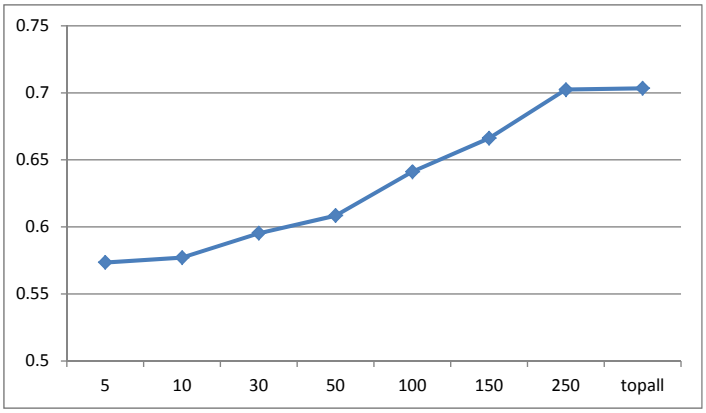

Figure 3.16: *The MAP of* Prototype-Set *for different numbers N of top images to be reranked.*

the Web Queries dataset demonstrate that the proposed method outperforms all the existing supervised and unsupervised reranking methods. It improves the performance by 25.48% over the text-based search result by combining prototypes and textual ranking features.

A natural extension of the approach described in this chapter would be to apply the proposed methods to learn concept models from image search engines in a semi-automatic fashion. Compared to the fully automatic methods [57], the semi-automatic approach could learn the concept models for any arbitrary concept much better and with only little human supervision.

While our proposed methods have proved effective for reranking image search results, we envision two directions for future work to further improve the reranking performance. First, we could further speed up the *Prototype-Set* method variant while decreasing the precision degradation. Since top images are incrementally added into the multiple-set prototypes to train the meta rerankers, one of the possible approaches in this direction is to utilize the online learning algorithms [52]. Second, although we assume that the rank position is generally correlated with the relevance value of the image found there, and while our results show that this assumption can be regarded valid in a general case, still deviations from this expectation can occur for individual queries. Hence, we could work on improving the proposed reranking model to make it more query-adaptive. One possible approach here would be to automatically estimate the query-relative reliability and accuracy of each meta-reranker and then incorporate it into the reranking model. Another approach may be to learn the reranking models for different query classes.

Chapter 4

Learning to Rerank Web Images: Reflections and Recommendations

[1]

This chapter reviews recent advancements in developing approaches to web image search reranking. A categorization of related theories and algorithms is provided, accompanied by a mathematical formulation, analysis and discussion per category. Limitations of the existing approaches are highlighted and recommendations are made on what we believe to be the most critical research directions to improve the efficiency, effectiveness and overall utility of web image search reranking technology.

[1]This chapter was published as: Linjun Yang, Alan Hanjalic, "Learning to Rerank Web Images," IEEE Multimedia Magazine, To Appear [129].

American Democrat officials today said Barack Obama is seriously considering Hillary Clinton or Bill Richardson to be his secretary of State. Richardson the New Mexico Governor has a fair bit of foreign policy under his belt with good qualifications these include a UN ambassadorship in the Clinton Administration plus diplomacy qualifications in the Sudan and North Korea. For what it is worth I reckon Hilary will be asked personally as after she stepped down in favour of B Obama she tirelessly worked her utmost for him; I could be wrong of course. More on that can be read here on the Sydney Morning Herald's online news. I have been receiving a lot of e-mail from friends lately who seem to wish to return to the Republican way and are back to their usual thing of sending mail to how it was prior to the election. I ignore it in the main but often wonder why they bother to do it, the man (yes he is the man now isn't he) got voted in fair and square by the American people he never squeaked in on a court decision like GW Bush did either he was actually voted into office. We here in Australia had one hell of a

Figure 4.1: *Illustration of the mismatch between an image and the surrounding text. This image (url:http://thewhizzer.blogspot.com/2008_11_01_archive.html) is returned by a popular web image search engine for the query "george w bush".*

4.1 Introduction

The existing web image search engines retrieve and rank images mostly based on matching the text queries with textual information accompanying the images on web sites, including the tags, comments, surrounding text, title, alt text and url. While the image retrieval performance can be good for many queries, the precision of the returned results is still not high in a general case. The major bottleneck is the likely mismatch between the image content and the text from the web pages, which is not always rightfully assumed to be associated with the image and to reveal precisely those aspects of the image content that are demanded by the query. This mismatch is illustrated by an example in Figure 4.1.

Image search reranking attempts to resolve this bottleneck by relying in the image search process not only on the text information channel, but also on the visual one. There, the ranked list of images obtained via search in the text channel

is considered as a noisy, but informative baseline. The visual content of the images is then deployed to reduce the ambiguity in the list and move more of the relevant images towards the top of the list. This process is illustrated in Figure 4.2.

Initially, the development of the image search reranking methods was based on the rationale that the consistency in the content of the relevant images should be observable in both the textual and visual domain. Enforcing the content consistency in both domains simultaneously has typically been attempted through an optimization approach, various realizations of which have led to many different proposals for image search reranking over the past several years [36][102]. Since the content consistency criteria in these methods have been defined largely ad hoc, recently proposed reranking methods have tried to incorporate sophisticated machine learning mechanisms into the process so that more reliable reranking models can be generated.

The goal of this chapter is to review the trends that have characterized the research on image search reranking, to discuss the problems found underway and to identify promising ideas that should guide future activities in this research direction. We will do this by performing a categorization of reranking theories and algorithms, which will be accompanied by a mathematical formulation, analysis and discussion per category. Finally, recommendations will be made on what we believe to be the most critical research topics to improve the efficiency, effectiveness and overall utility of web image search reranking technology. In view of the fact that these topics not only address the scientific concepts, but also the issues related to optimization of the implementation of reranking mechanisms in real-life search engines, this chapter not only targets researchers, but also web system architects and search engine developers.

We start in Section 4.2 with a general mathematical formulation of the image search reranking problem. This is followed in Section 4.3 by a categorization, an overview, and a discussion of the existing solutions to this problem. Section 4.4 highlights the issues that we consider important for the future research on image search reranking. Section 4.5 concludes the chapter with a summary of recommendations for future work.

4.2 Problem formulation

In the following we will give several definitions to mathematically formulate the image search reranking problem.

Definition 4.2.1. *A ranking score list* $\mathbf{r} = [r_1, r_2, \cdots, r_N]^T$*, is a vector of real numbers, each of which corresponds to the ranking score of an image in the image set* $\mathcal{D} = \{d_1, d_2, \cdots, d_N\}$*.*

Definition 4.2.2. *A ranking list* $\mathbf{l}$ *is a permutation of* $\mathcal{D}$ *sorted by the ranking scores in descending order.*

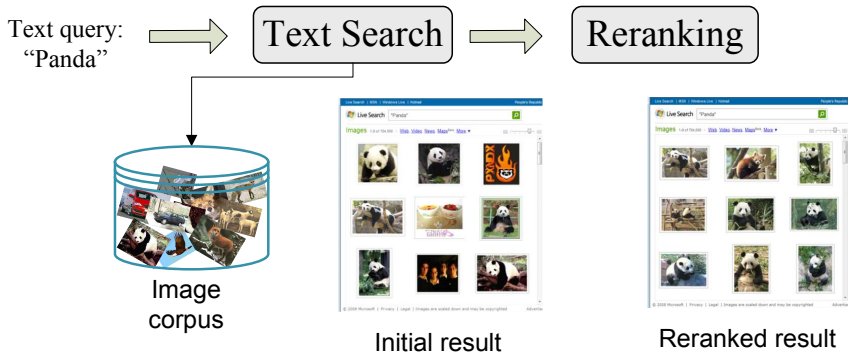

Figure 4.2: *Illustration of the image search reranking process.*

In general reranking can be regarded as a mapping from the initial ranking list to the desired (target) ranking list. This mapping is generated using a *reranking model* that recomputes the ranking score list based on the available additional (e.g. visual) information.

Definition 4.2.3. *A reranking model is defined as a function*

$$r_j^i = f(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i). \tag{4.1}$$

Here, $\mathcal{D}^i$ is a collection of images d_j^i returned for the query q^i, which may be represented by features from different modalities such as text and visual. Furthermore, $\bar{\mathbf{r}}^i$ is the ranking score list of images $\mathcal{D}^i$ in the initial search result, while r_j^i is the final ranking score for image d_j^i.

The existing reranking methods differ from each other mainly in the way they derive the reranking model f. In the following, we will propose a categorization of existing image search reranking methods and discuss how the reranking model can be derived per category.

4.3 Categorization and analysis of approaches

The reranking function f can be written in general as

$$f = h \circ g(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i), \tag{4.2}$$

where $\circ$ represents the composition of two functions, a *query-independent function* h and a *query-dependent function* g. The query-independent function keeps a unified representation with identical parameters across different queries, while the query-dependent function is adjustable for each query based on the information derived from the initial ranking.

In many reranking approaches [121][37][38][43][102], h is a pre-defined query-independent function or even a constant function and g a scalar function learned from the initial ranking returned by the search engine. We can therefore say that in this case the (parametric or non-parametric) reranking model is generated through *learning from search engine*. Since function g is dependent on the query for which the initial ranking was generated, these approaches can be said to deploy *query-dependent learning*. Learning of g here is typically unsupervised, since the samples used for learning are labeled by the search engine based on the initial ranking and are not provided by a human. Consequently, the approaches from this category are also referred to as *unsupervised reranking*. The features employed in such approaches to construct image similarity models underlying the function g are usually visual features such as color, texture, and local gradient-based features, which are only related to the content of individual images and therefore said to be *query-independent*. We analyze and discuss these approaches in Section 4.3.1.

In another category of approaches [126][50], function g is defined as a function to compute a vector of reranking features from the initial ranking, and h is learned from human-labeled data. The reranking features produced by g embed the information about the relevance of an image to a query, and can be characterized as *query-dependent*. As model learning has shifted to function h, we can say that the reranking model is generated through *learning from human supervision*. These approaches have built on the success of the learning-to-rank [59] concept that was introduced in the field of information retrieval. In these methods, human supervision is deployed to learn more sophisticated image relevance criteria and develop a better reranking model than the ad-hoc one learned in an unsupervised fashion from the search engine as in the first category of approaches described above. A detailed analysis and discussion of the approaches falling into this category can be found in Section 4.3.2.

A third category of approaches combines the advantages of the above two paradigms and produces a reranking model through *learning from search engine and human supervision*. The approaches falling into this category are discussed in Section 4.3.3.

4.3.1 Learning from search engine

The approaches of learning from search engine are mostly based on two underlying assumptions, the *pseudo-relevance feedback (PRF)* assumption and the *visual consistency* assumption.

According to the PRF assumption, the top-ranked images in the text-based search result can be considered relevant. Then a ranking model can be learned from these (pseudo-relevant) images and deployed to predict the refined ranking scores. Taking the method [121] as an example, where the PRF assumption is implemented by means of a SVM (Support Vector Machine) classifier [109], the

reranking model in this case can mathematically be defined as follows:

$$\begin{aligned} f &= h \circ g(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i) \\ &= g(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i) \\ &= \hat{\mathbf{w}} d_j^i, \end{aligned} \tag{4.3}$$

with

$$\begin{aligned} \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C_1^+ \sum_{d_j^i \in \mathcal{D}^{i+}} \max(0, 1 - \mathbf{w}^T d_j^i) \\ + C_1^- \sum_{d_j^i \in \mathcal{D}^{i-}} \max(0, 1 + \mathbf{w}^T d_j^i). \end{aligned} \tag{4.4}$$

The reranking model f in (4.3) is an implicit function mapping the initial text-based result to the refined ranking scores. It is an ad-hoc function designed by domain experts and based on the hypothesis that the reranking criterion can be derived from the patterns discriminating the top-ranked results from the rest in the initial list. Here $\mathcal{D}^{i+}$ is the collection of pseudo-relevant images corresponding to the top-M images in the text-based search result and $\mathcal{D}^{i-}$ is the collection of pseudo-irrelevant images which are sampled from the bottom of the text-based search result or from the entire collection [121]. Alternatively, the set $\mathcal{D}^{i+}$ could also be constructed from the click-through log of an image search engine [40]. The refined query model $\hat{\mathbf{w}}$, which is an intermediate variable of the reranking model f, is learned from the initial text-based result and is query-dependent. In this sense, it can be said that the reranking model (4.3) resembles the query-dependent function g, while the query-independent function h can be seen as a non-informative constant. Finally, C_1^+ and C_1^- are the parameters to control the tradeoff between the regularization and the loss from positive and negative samples.

According to the visual consistency assumption, the visually consistent images should be ranked close to each other. We illustrate the possibilities for implementing this assumption on the example of the Bayesian reranking approach [102] that also more explicitly reveals the trade-off between this assumption and the assumption common to reranking in general, namely that the bias towards the noisy but still informative initial results list should be preserved:

$$\begin{aligned} &f = h \circ g(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i) = \hat{\mathbf{r}}_j^i, \\ &\hat{\mathbf{r}}^i = \arg\min_{\mathbf{r}^i} \ \frac{1}{2} \sum\nolimits_{j,k} v_{jk}^i (r_j^i - r_k^i)^2 + C_2 \sum (1 - \frac{r_j^i - r_k^i}{\bar{r}_j^i - \bar{r}_k^i})^2. \end{aligned} \tag{4.5}$$

Here, v_{jk}^i is the visual similarity in terms of the feature vector between the images d_j^i and d_k^i ranked at the positions j and k in the initial list obtained for query q^i, and C_2 is the trade-off parameter. The visual similarity function v_{jk}^i here is defined a priori and does not change across queries. It can therefore be said

to resemble the query-independent function h, which in this case serves as an argument of function g.

If we compare the methods used in the examples above, we can say that in the case of Bayesian reranking, the reranking model f in Eqn. (4.5) is also an implicit expert-designed function mapping the initial text-based search result to the new refined ranking. However, different from the model in Eqn. (4.3), we do not need to explicitly infer a query model, but only to estimate the new ranking scores for all candidate images.

While having the advantage of not requiring human supervision and therefore to scale well across a broad range of different queries, this category of reranking approaches suffers from insufficient reliability of the assumptions under which the initial text-based image search result is employed in the reranking process and from a missing link between these assumptions and the human notion of relevance of retrieved images. Specifically regarding the examples discussed above, the PRF assumption may not be satisfied well by the existing image search engines. Furthermore, the visual consistency assumption steers towards optimizing the results list to be visually consistent, which does not necessarily guarantee semantic relatedness between these images and the query and therefore the actual relevance for the users.

4.3.2 Learning from human supervision

The reranking approaches discussed in this section learn the reranking model from human-labeled samples and are therefore referred to as *supervised reranking* [126][50]. We adopt here a definition of supervised reranking as formulated in [126].

Definition 4.3.1. *Supervised reranking is defined as a process of learning a reranking model f from the given training samples $\{\mathcal{D}^i, \bar{\mathbf{r}}^i, q^i, \mathbf{r}^i\}$, where $\mathcal{D}^i, \bar{\mathbf{r}}^i, \mathbf{r}^i$ are the initially returned documents, initial ranking and the ground-truth ranking corresponding to the query q^i. The learning process can be formulated as the process of minimizing the loss function*

$$f^* = \arg\min_f \sum\nolimits_i \triangle(\mathbf{r}^i, [f(d^i_j, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i)]_{j=1}^{N^i}). \tag{4.6}$$

where $\triangle$ measures the loss between the ground-truth $\mathbf{r}^i$ and the prediction $\hat{\mathbf{r}}^i = [f(d^i_j, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i)]_{j=1}^{N^i}$. N^i is the number of images to be reranked.

This definition indicates that the reranking model f in supervised reranking is not a pre-defined function as in the unsupervised reranking case, but is estimated by optimizing the loss function $\triangle$. The loss function is defined on a few manually labeled samples of selected queries and images in order to measure whether the predicted ranking from the learned model is in accordance with the manual labels.

Since it is impossible to learn a reranking model for each possible query, a general model applicable to all queries should be learned. In order to achieve this,

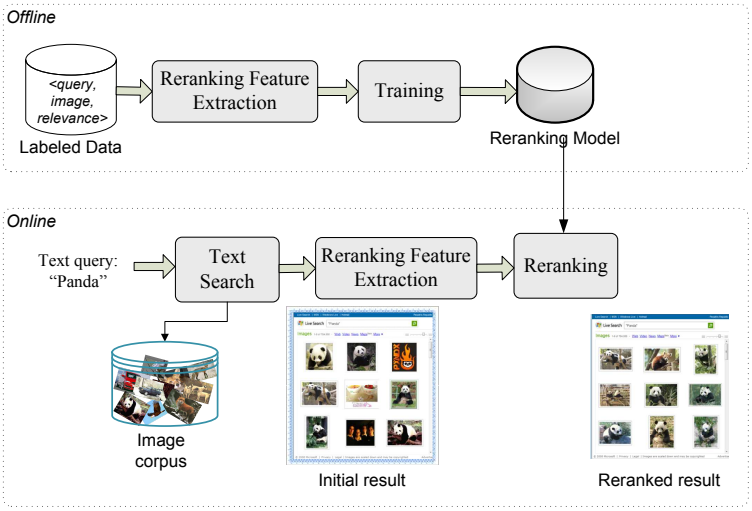

Figure 4.3: *An illustration of the supervised reranking process (adopted from [126]).*

the query-independent component of the reranking model f needs to be developed that is learned offline using several representative queries, but that can model the patterns common across queries. Figure 4.3 illustrates the combination of the query-independent and query-dependent components in a supervised reranking approach.

A supervised reranking approach can be developed by representing f as a weighted combination of multiple terms, where each of the terms describes one aspect of the relevance between a query and an image where a query is usually represented using its text-based result returned by the search engine and where the weights to combine the components are common and can be learned across queries:

$$\begin{aligned} f &= h \circ g(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i) \\ &= h(g(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i)), \end{aligned} \tag{4.7}$$

with

$$h = \sum_k \mathbf{u}_k g_k \tag{4.8}$$

and

$$g = [g_1(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i), ..., g_k(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i), ...]. \tag{4.9}$$

Here we refer to g_k as reranking features realized in the form of meta rerankers. In this way, the process of learning the query-independent reranking models $\mathbf{u}_k$ can be reduced to a standard learning-to-rank problem [59] that requires a few labeled samples. Hence, the existing learning-to-rank algorithms including Ranking SVM [44] can be deployed to learn the weights $\mathbf{u}_k$.

Table 4.1: *An overview of the eleven reranking features proposed in [126].*

HV_N	Hard Voting of Neighbors
RSV_N	Initial Rank based Soft Voting of Neighbors
$NRSV_N$	Neighbor Rank Weighted Initial Rank based Soft Voting of Neighbors
HV_R	Hard Voting of Reciprocal Neighbors
RSV_R	Initial Rank based Soft Voting of Reciprocal Neighbors
NSV_R	Neighbor Rank based Soft Voting of Reciprocal Neighbors
$NRSV_R$	Neighbor Rank Weighted Initial Rank based Soft Voting of Reciprocal Neighbors
PRF_d	Local Density Estimation for PRF
PRF_{dv}	Duplicate Voting for PRF
PRF_{sdv}	Soft Duplicate Voting for PRF
IR	Initial Ranking

Effective embedding of query information into the meta rerankers is critical in order to be able to learn $\mathbf{u}_k$ and to make it generalize across queries. In [126], eleven reranking features were proposed, as listed in Table 4.1. These features are derived from the initial text-based ranking and from the visual content analysis of the initially returned top images. As can be seen from Table 4.1, the features are mostly based on simple counting, like the number of near-duplicates in top results or the weighted number of visual neighbors of each image, and are therefore computationally lightweight. Their effectiveness has been demonstrated in [126] on a moderate dataset collected by issuing 29 queries on three image search engines including Bing, Google, and Yahoo! .

Krapac et al. [50] developed another strategy to build the meta rerankers. First, the bag-of-words feature extraction method is applied to top images in the initial list. Visual words are then aggregated over all top images and sorted according to their word frequencies to serve as the visual surrogate of the textual query. Finally, for each image in the initial list, the reranking feature vector is extracted by comparing the frequencies of each visual word in the image to that in the visual query surrogate to form a binary vector of visual word presence.

Just like in the case of unsupervised reranking, the examples of approaches discussed above indicate that the reranking features are here again based on the visual consistency assumption and pseudo-relevance feedback assumption. While these assumptions are unreliable if serving alone as the basis for building a reranking model, supervised reranking improves the reliability by steering the model learning using human-labeled samples. The critical aspect here is, however, that the learning process is designed such that human supervision does not reduce the scalability of the reranking approach in terms of the coverage of the query space.

4.3.3 Learning from search engine and human supervision

One of the drawbacks of the supervised reranking approaches is that the reranking features or meta rerankers are hand-designed by domain experts. In general, this may be insufficiently effective for discovering sophisticated information contained in the data that could be beneficial for reranking. Furthermore, due to large variations in text-based search results, it is virtually impossible to hand-design these features such that a reranking approach could work well on all search engines and for all queries. To address this disadvantage, automatic approach is required to learn the reranking features from the initial text-based result that are adaptive enough to grasp the characteristics of the underlying text-based image search engine.

In [127], a two-stage learning framework for image search reranking was proposed. In the first step, meta rerankers are learned automatically steered by the initial text-based ranking result. Then, in a second step, a reranking model is learned using a number of selected queries as discussed in the previous section. In this way, two categories of approaches, learning from search engine and learning from human supervision, are leveraged together to improve the quality of the reranking scores of the initial image search result.

The underlying assumption of the feature-learning step in [127] is that the images at different positions in the initial text-based ranking should be considered with different confidence values when acting as positive samples. These confidence values should also be roughly consistent across queries for a given image search engine. This assumption is reasonable since the modern image search engines optimize the relevance probability of top-ranked images, as evidenced by the widely adopted evaluation measures like Mean Average Precision (MAP) or Normalized Discounted Cumulative Gain (NDCG) that build on such optimization.

Figure 4.4 illustrates the basic idea of the reranking approach described above that consists of two learning stages. Multiple meta rerankers are built by employing different sets of images as positive samples. More specifically, top k images from the initial ranked list are used to learn the k^{th} meta reranker. The higher the image rank, the more meta rerankers it will be included in. In this way, the prior information that the top images have higher relevance probability can be implicitly incorporated in the process of learning the reranking function. However, the relevance confidences of images at different positions can still be adjusted based on human supervision. This is done by learning the combination weights of different meta rerankers from human-labeled data.

To illustrate the relation between the reranking approaches belonging to this category and the approaches discussed in the previous sections, we take the method [127] as an example. Mathematical formulation of the reranking model there adopts the expressions in Eqn. (4.7) and Eqn. (4.8), with the difference that g_k now represents the k^{th} meta reranker computed using a slightly modified query model from Section 4.3.1:

$$g_k(d_j^i, \mathcal{D}^i, \bar{\mathbf{r}}^i, q^i) = \hat{\mathbf{w}} d_j^i, \tag{4.10}$$

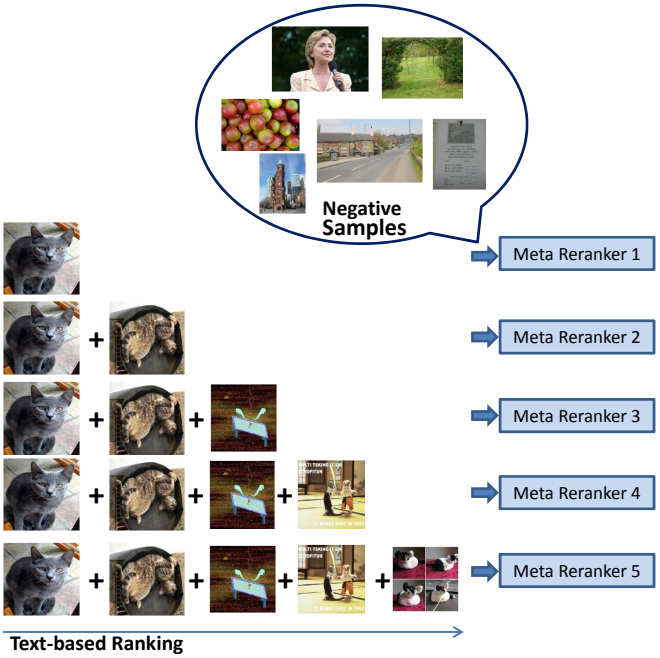

Figure 4.4: *An illustration of an approach to learning meta rerankers.*

with

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C_1^+ \sum_{d_j^i \in \mathcal{D}_k^{i+}} \max(0, 1 - \mathbf{w}^T d_j^i) + C_1^- \sum_{d_j^i \in \mathcal{D}_k^{i-}} \max(0, 1 + \mathbf{w}^T d_j^i). \tag{4.11}$$

Here $\mathcal{D}_k^{i+}$ is the collection of top k images in the initial text-based result, and $\mathcal{D}_k^{i-}$ comprises the negative samples selected from the background, e.g., sampled from the entire collection.

The assumption underlying the model (4.10) is more relaxed than the PRF assumption in that top-ranked results are not automatically assumed relevant, but the relevance probabilities of images are learned based on human-labeled training samples. We can see that SVM-based PRF in Section 4.3.1 is a special case of the above formulation, if the weight vector $\mathbf{u}$ is not learned from human labeled data but set to [0 0 ... 0 1]. Experiments reported in [127] and performed on the Web Queries dataset[2] have indicated that this approach outperforms the initial text-based by 23.9% and performs significantly better than many representative existing reranking approaches including PRF [121], Bayesian Reranking [102],

[2]http://lear.inrialpes.fr/~krapac/webqueries/webqueries.html

Table 4.2: *Performance comparison of various reranking methods.*

Methods	MAP	NDCG@10	NDCG@40
Search engine	0.569	0.682	0.633
PRF [121]	0.658 (+15.64%)	0.772 (+13.20%)	0.718 (+13.43%)
Bayesian [102]	0.643 (+13.01%)	0.766 (+12.32%)	0.709 (+12.01%)
Supervised-reranking [126]	0.665 (+16.87%)	0.769 (+12.76%)	0.733 (+15.80%)
Query-relative [50]	0.666 (+17.05%)	0.768 (+12.61%)	0.729 (+15.17%)
Two-stage learning [127]	0.705 (+23.90%)	0.828 (+21.41%)	0.778 (+22.91%)

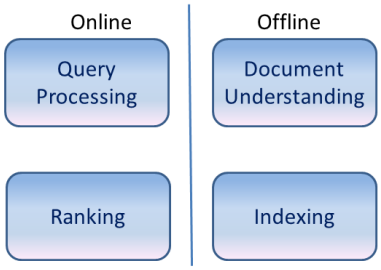

Figure 4.5: *The main components of a web image search engine.*

query-relative learning [50], and supervised reranking [126], which shows the effectiveness of the combined learning paradigm. The numerical results of various representative reranking approaches on the Web Queries dataset, as adopted from [126], are shown in Table 4.2.

4.4 Remaining challenges

While significant progress has been achieved in web image search reranking over the past years, we point in this section to a number of critical limitations of the existing reranking approaches and recommend promising research directions towards addressing these limitations and improving the overall utility of web image search reranking solutions. These issues include image search system architecture, search results diversification, query adaptivity of the reranking mechanism, and maximizing the benefit from human supervision.

4.4.1 System architecture

Although image search reranking can be seen as a post-retrieval ranking refinement step, its successful deployment poses substantial requirements on the system architecture of a web image search engine. As illustrated in Figure 4.5, such engine typically consists of four main components:

- *query processing*, where user-generated queries are transformed into a format interpretable by the engine, which includes query alteration and expansion;
- *document understanding*, which extracts meta data from the associated web page for representing an image;
- *indexing*, which provides an efficient organization of the images' metadata to speed up the retrieval; and
- *ranking*, which retrieves and ranks images based on the relevance of images to the query.

The reranking step is mainly related to the latter three components and imposes their adjustments at the system architecture level. The document understanding component needs to be modified so that the visual features can be extracted from the images after images have been collected. In order to keep the throughput of image crawling and processing acceptable, the time cost for the visual feature extraction needs to be limited to the order of milliseconds. This imposes critical challenges on the development of the extractors of visual features typically deployed in a reranking context, such as the pyramid histogram of visual words (PHOW) [13].

Since the dimensionality of the visual features is usually high, a large additional memory space is required to store them. For example, storing the PHOW features for one million images using naive approaches may require several gigabytes of memory, which would double the memory cost of the currently deployed image search engines. Reducing the memory requirements of the visual features while maintaining their effectiveness for reranking is another critical challenge posed on the development of future web image search reranking solutions.

Finally, a third architecture-related challenge is posed by the fact that the time cost of visual reranking in the ranking component should be in the order of milliseconds in order to be able to maintain the current query response speed. This, however, is difficult to achieve due to several complex steps of a general image search reranking algorithm, including the distance computation, model training and ranking score computation.

4.4.2 Diversification

The objective of most of the existing visual reranking approaches is to optimize the relevance of the image search result, for the purpose of which the measures, such as MAP or NDCG, are deployed. Other criteria imposed by the users, such as the diversity of the retrieved images have, however, not be taken into account to a significant extent. A possible reason for this is a high difficulty of capturing the diversity in an objective measure. The measures typically deployed rely mainly on visual diversity (e.g. [108]) and may not be powerful enough to capture the semantic diversity of the retrieved visual content. Without the semantic diversity

being taken into account, the retrieved images may contain trivial samples not being capable of satisfying the information need of the user. The problem of semantic diversification also increases through the fact that some of the popular reranking approaches are based on the assumptions (e.g. the visual consistency assumption) that conflict with the diversification criteria. This poses a challenge on the development of future reranking mechanisms that should jointly optimize both the relevance and diversity of the reranked image search results.

4.4.3 Adaptivity to a query

Due to a large variance among queries, one unified reranking model for all queries is not likely to be optimal across the entire query space. Indeed, while image search reranking has been shown to bring significant improvement of search results for many queries, it still degrades the performance of the initial search results for some of the queries. Although this degradation is statistically much less frequent than the achieved performance improvement, the main practical problem related to it is that degradation is difficult to predict, which may have negative impact on the user experience, in particular if the user prefers the queries for which reranking does not perform well. Essentially, to solve this problem an analysis needs to be performed in order to identify those queries that would benefit from reranking and under which conditions. We refer to this process as *rerankability analysis* and consider it one of the most important challenges that need to be pursued in order to bring the image search reranking technology to the sufficient utility level.

An important issue that would need to be considered during rerankability analysis are the visual features used to rerank the results list for a given query. For example, the color features being important for the query "red apple" would clearly not be that suitable for the query "street view", where GIST feature [74] may be much more powerful in describing the scene appearance. Since there is no single feature which can perform best for all queries, one of the challenges underlying the development of rerankability analysis methods is to create a reliable feature space from which optimal feature selection can be drawn for a given query space.

Once the feature space has been defined, methods need to be found to perform the actual selection of the most suitable reranking option for a given query. Preliminary results in this direction were reported by Tian et. al. [100]. There, a method was proposed for selecting the best performing ranking option taking as input a set of ranking lists generated by a text-based search baseline or a number of reranking methods deploying different visual features. While in [100] reranking selection was addressed by a preference learning model operating on carefully designed features extracted from different ranking lists, also an approach based on the idea of coherence-based query performance prediction (QPP) [84] could be deployed for this purpose. While the results reported in [100] and [84] were promising, still a substantial body of new research is needed to improve the efficiency and effectiveness of the post-retrieval list selection methods, but also to

explore more direct ways of evaluating different reranking options for a given query and the initial results list.

4.4.4 Learning from search engine with light supervision

The approach of learning from both the search engine and human supervision as described in Section 4.3.3 currently represents the most promising direction of developing web image search reranking methods. This, however, is still not the optimal solution since the learning process is split into two separate steps, which does not allow cross-usage of the information contained in the initial text-based result and the information derived from the human supervision between the two steps. We therefore envision a framework that is to be explored in our future work and that will make a more effective use of the information available at different stages of the reranking process, for instance by employing human supervision to steer the process of learning from search engine. The challenge here is to maximally benefit from human supervision, but without jeopardizing scalability.

4.5 Conclusions and recommendations

In this chapter we reviewed recent advancements in web image search reranking. We grouped the existing approaches in three categories and used this categorization to depict the main characteristics of the development of reranking methodologies over the past years. Advantages and disadvantages of the approaches per category were highlighted, which led to an overview of the main challenges we see in front of the research community addressing the improvement of the efficiency, effectiveness and overall utility of web image search reranking technology in the future.

In addition to the challenges related to system architecture (computational efficiency, compactness of visual feature representation), expansion of the reranking criteria (from relevance only towards combined relevance and diversification) and adaptivity of the reranking mechanism to the query (rerankability analysis), we also see a high importance in further improving the reranking approaches by maximizing the benefit from human supervision, as envisioned in Section 4.4.4.

Another direction for future work would be an in-depth study of various aspects of the learning process underlying the development of reranking models in the approaches discussed in this chapter. Such analysis should optimally map this process onto the most efficient and effective recently proposed learning-to-rank algorithms [59]. Specifically for the methods deploying learning from human supervision, it should be investigated how many queries are sufficient for training a reasonably good reranking model and how to select informative queries and images for human labeling to construct the training set for learning an effective reranking model.

Part II

Leveraging Context for Example-based Image Search

Chapter 5

Object Retrieval using Visual Query Context

[1]

In this chapter we address the problem of object-based image retrieval, here referred to simply as object retrieval. Object retrieval aims at retrieving images containing objects similar to the query object captured in the region of interest (ROI) of the query image. While existing object retrieval methods perform well in many cases, they may fail to return satisfactory results if the ROI specified by the user is inaccurate or if the object captured there is too small to be represented using discriminative features and consequently to be matched with similar objects in the image collection. In order to improve the object retrieval performance also in these difficult cases, we propose in this chapter an object retrieval method that exploits the information about the visual context of the query object and employ it to compensate for possible uncertainty in feature-based query object representation. Contextual information is drawn from the visual elements surrounding the query object in the query image.

[1] This chapter was published as: Linjun Yang, Bo Geng, Yang Cai, Alan Hanjalic, Xian-Sheng Hua, "Object Retrieval using Visual Query Context." IEEE Transactions on Multimedia vol.13, no.6, pp.1295-1307, Dec. 2011 [123].

Figure 5.1: *Illustration of the case where the bounding box does not accurately represent the search intent. In addition to relevant objects (the Pitt Rivers in Oxford), the bounding box includes some non-relevant objects as well as some parts of the background.*

5.1 Introduction

Recent advances in computer vision, and in particular in the development and deployment of local visual feature descriptors, like SIFT [61] have boosted the popularity of object retrieval and its adoption in real-life applications and products. For example, TinEye [7] released an application based on near-duplicate web image search, while Google Goggles [5] allows users to take a picture using a mobile phone and then to retrieve information related to the object in the picture.

In a typical object retrieval system, a user first selects an example (query) image and then draws a bounding box in that image around the object of interest to specify the search intent. Local features are then extracted from the bounding box and then quantized into the so-called visual words. This "bag of visual words" representation is used to match relevant images in a collection, where the relevance is often computed by proven techniques in the field of information retrieval. For example, Philbin et al. [77] employed the cosine retrieval model based on *tf-idf* vector representation of images and Geng et al. [31] studied the methods based on language modeling [137] for the purpose of object retrieval.

While object retrieval based on the visual words is generally effective, it may not achieve a reliable search result in cases where the visual words extracted only from the bounding box are unable to reliably reveal the search intent of the user. Firstly, the bounding box is typically only a rough approximation of the ROI (region of interest) representing the query object. For example, as shown in Fig. 5.1, the bounding box may not represent the ROI accurately since it is simply a rectangle while the ROI has a complex shape. Therefore, the visual words extracted from the bounding box may also carry information that is unrelated to

Figure 5.2: *Since the bounding box is too small, the search intent cannot be estimated reliably, which results in non-relevant search results like the image containing people.*

the search intent. Secondly, in some cases where the ROI is too small, or where the visual content within the ROI lacks texture, the number of visual words in the bounding box may be insufficient to perform a reliable relevance estimation, with the consequence that irrelevant images may be returned. For example, a small bounding box used in the example in Fig. 5.2 resulted in retrieving an image of people for the query image showing a building.

The two challenges mentioned above essentially attribute to the uncertainty of the information contained in the bounding box to reflect the true search intent of the user. This uncertainty could be handled by taking into account the fact that objects in real-life images hardly occur in isolation [30][79][75][104]. In this sense, the visual information outside the ROI can be seen as the context in which ROI is specified as a search query. By combining the information from the ROI and from the context, a better query representation could be obtained. For example, as shown in Fig. 5.3 the water and the lotus leaves surrounding the ROI (lotus flower) specified by the bounding box can help estimate a better representation for the query "lotus", especially when the visual word representation leads to poor search results for flowers, while it performs relatively well for leaves.

In this work we follow the rationale described above and propose a contextual object retrieval (COR) model that effectively employs the visual context information together with the ROI to improve object retrieval in general, specifically in the difficult cases discussed earlier in this section. Following the common practice in the object retrieval field (e.g [31]), we base our model on the language modeling approaches for information retrieval [137]. Different from the conventional methods, which estimate the query language model only based on the visual words within the bounding box, our proposed query language model is estimated using the visual words from both the ROI and the visual context. These visual words are weighted using the search intent scores that are based on the uncertain observation of the search intent, i.e., the bounding box, and the prior information derived from a saliency map of the image.

We evaluated the proposed model experimentally on three representative datasets. One of them is the publicly available Oxford building dataset comprising 5K images of Oxford landmarks [6][77]. To test the performance of our method in a large-scale image retrieval setting, we constructed a second dataset by combining the 5K images from the Oxford building set and 500K images from ImageNet

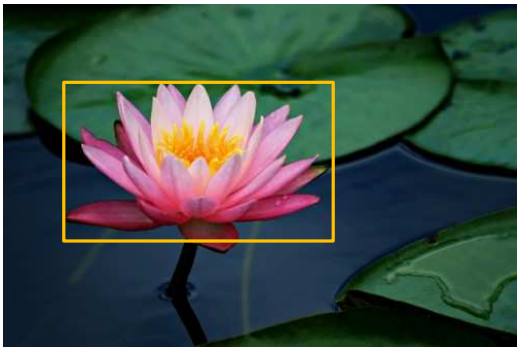

Figure 5.3: *Illustration of the usefulness of the visual context in object retrieval. The search intent (lotus flower) within the bounding box can be expressed more precisely by also taking into account the context represented by the water and lotus leaves.*

[25]. Furthermore, to investigate the usefulness of the visual context and the effectiveness of the proposed model on a broader range of search intent categories, we collected a dataset by crawling 1 million images from Internet and generating query objects of landmarks, books, logos, paintings, animals, etc.

Although, intuitively, the introduction of context consisting of objects and background that are irrelevant to the query may be said to introduce noise in the retrieval process, our experimental evaluation demonstrated that the benefit of including the context in the retrieval model in the way introduced in this chapter is stronger than the noise this context may introduce. Furthermore, the evaluation showed that the proposed COR model outperforms the alternative object retrieval approaches.

After reviewing the related work in Section 5.2 and positioning the contribution of this chapter with respect to it, we introduce our COR model in Section 5.3. In Section 5.4, we propose two methods for computing the search intent score. The results of the experimental evaluation are reported and discussed in Section 5.5, while Section 5.6 brings the most important conclusions and recommendations for future work.

5.2 Related Work and Contribution

User-generated queries are often simple and cannot contain sufficient information to properly reveal the user's search intent. To deal with this problem of query uncertainty or query ambiguity, several general approaches have been proposed in the past, the most well-known of which is relevance feedback [142]. In this approach, the feedback from users regarding the relevance of the search results in the initial search step is exploited in an iterative procedure to estimate a better retrieval model. Although involving the user in the retrieval loop is conceptually

promising for learning the search intent, practical issues related to the effort required from the user in multiple iterations of the learning process have prevented the adoption of this idea in real-life web search applications [67]. In this chapter, we aim at refining the query model through a one-step interaction, in which the user is only required to specify the ROI in the query image. We then exploit the information from the ROI and the context in which ROI is placed to learn the query model.

Context has already established itself as an important auxiliary information source for developing robust and reliable image retrieval solutions [63] and improving information retrieval in general [9][15][12]. Contextual information derived from the user profile, from a user's search interaction, or from a user's everyday activity can help infer user's search preferences and intent and in this way disambiguate the query [55][135], while the text surrounding an image on a web page or the tags (including GPS), comments, and other types of metadata (e.g., exposure time, ISO, etc.) attached to an image can help learn more about the relevance of that image with respect to the user's information need [47][91][130].

In this chapter we focus on the local visual query context, which is, due to its direct availability, particularly suitable for enriching the information captured in the bounding box and in this way for reducing the gap between the ROI estimation and the user's search intent. The visual context has been widely exploited in the object recognition tasks in the fields of human vision and computer vision [30][104][75]. It proved to be useful for disambiguating visual objects that are cluttered, blurred, or with unfamiliar appearances to a human or computer vision system. In such cases, the visual context can be leveraged to help improve the reliability of the object recognition.

While the state-of-the-art contextual object recognition methods largely exploit the visual context that comprises the global features of the image containing the object [104][75], the method proposed in this chapter relies on the visual context derived from weighted visual words and is suited for modern image retrieval approaches utilizing image representations based on the bag-of-visual-words concept [93][31]. In such approaches, the SIFT features [61] are first extracted from the images and then quantized into visual words. Then the visual words can be indexed using inverted file system [67]. Finally, the cosine retrieval model [77] or language model [31] can be adopted to rank images according to their relevance with respect to the query.

In our COR model, the visual context is employed to estimate the search intent score for each visual word in the query image. This score indicates how likely the image region represented by the visual word reflects the search intent of the user. We estimate this score using two approaches, one being based on the distance of image pixels from the bounding box and the other one being based on the color coherence of the pixels. We refer to these two methods as spatial propagation and appearance propagation, respectively. In the first case, we employ the dual-sigmoid function fitting to compute the scores, while in the second one matting algorithms [83][112] are used.

5.3 Contextual Object Retrieval Model

We approach the development of our COR model by first defining the main terms and notations we are going to use in the process. Then, we briefly introduce the basic language model approach that our COR model is based on, namely the well-known Kullback-Leibler divergence retrieval model, which will then be expanded to become context-aware by taking into account the search intent scores. The strategies for computing these scores are explained in detail in Section 5.4.

5.3.1 Definitions of basic terms

The task addressed in this chapter is to develop a mechanism for returning a ranked list of images relevant to the information need represented by the query. The essential part of this task is to compute the relevance of the image $\mathbf{d}$ in the database with respect to the query $\mathbf{q} = \{\mathbf{q}^I, \mathbf{q}^b\}$, which can be said to formally consist of the example image $\mathbf{q}^I$ and the bounding box specification $\mathbf{q}^b = [x_l, y_l, x_r, y_r]$. Here, (x_l, y_l) and (x_r, y_r) are the coordinates of the top left and bottom right point of the rectangle, respectively.

The query image and the images in the database are represented as a sequence of visual words using the following procedure. Firstly the interest points are detected in the images based on methods like DoG (Difference of Gaussian) [61] or Harris Affine detectors [71]. Then for each of the detected interest points SIFT (Scale Invariant Feature Transform) descriptors [61] are extracted to represent the local region around each interest point. The SIFT descriptors are then quantized into the so-called visual words using the K-means vector quantization method [77]. As a result, the query image is represented as $\mathbf{q}^I = [(q_i, p_i)]_{i=1}^{M_q}$ and the images in the database, further referred to as *documents*, are represented as $\mathbf{d} = [d_i]_{i=1}^{M_d}$. Here, q_i and d_i are the extracted visual words from the query and a document, respectively, p_i is the corresponding position of a visual word in an image, and M_q and M_d are the numbers of visual words in the query and database images, respectively. For the purpose of general explanations of terms and models, we also employ w or w_i to denote a visual word in an arbitrary image.

5.3.2 Kullback-Leibler divergence retrieval model

Once the images are represented as sets of visual words, classical information retrieval models can be employed directly for image search. Among such models, the language modeling approach has been one of the most popular approaches due to its sound theoretical foundation and flexibility to introduce additional components, such as relevance feedback [137]. In the language modeling approach for information retrieval, a language model, usually the unigram model $p(w|\mathbf{d})$, is estimated for the words w for each of the documents $\mathbf{d}$ in the database. Then the relevance between the query and the document is estimated as the query likelihood given the document. Using the visual word notation we introduced

before for our specific image retrieval context, this likelihood can be written as

$$p(\mathbf{q}|\mathbf{d}) = \prod_{i=1}^{M_q} p(q_i|\mathbf{d}). \tag{5.1}$$

In [51], Lafferty and Zhai further generalized the language modeling approach as a risk minimization problem. The risk of returning a document $\mathbf{d}$ given the query $\mathbf{q}$ is defined as

$$\begin{aligned} R(\mathbf{d};\mathbf{q}) =& R(a = \mathbf{d}|\mathbf{q}, \mathcal{C}) \\ =& \sum_{r\in\{0,1\}} \int_{\theta_Q} \int_{\theta_D} L(\theta_Q, \theta_D, r) \times p(\theta_Q|\mathbf{q}) \\ & p(\theta_D|\mathbf{d}) p(r|\theta_Q, \theta_D) d\theta_Q d\theta_D, \end{aligned} \tag{5.2}$$

where $a = \mathbf{d}$ is the action to return the document $\mathbf{d}$ for the query $\mathbf{q}$, $\mathcal{C}$ is the collection of documents in the database, r indicates the relevance of the document $\mathbf{d}$ to the query $\mathbf{q}$, and where θ_Q and θ_D are the language models for the query and the document, further refereed to as the *query model* and *document model*, respectively. Finally, L is the loss function, which can best be modeled using the Kullback-Leibler (KL) divergence between the query model and a document model [51]. By using the KL divergence to estimate the loss function the risk function can be formulated as

$$R(\mathbf{d};\mathbf{q}) \propto -\sum_{w_i} p(w_i|\hat{\theta}_Q) \log p(w_i|\hat{\theta}_D) + \xi_q, \tag{5.3}$$

where

$$\begin{aligned} \hat{\theta}_Q =& \arg\max_{\theta_Q} p(\theta_Q|\mathbf{q}) \\ \hat{\theta}_D =& \arg\max_{\theta_D} p(\theta_D|\mathbf{d}) \end{aligned} \tag{5.4}$$

are the maximum a posteriori estimations of the query and document models. The term ξ_q is a query-dependent constant and can therefore be ignored when Eqn. (5.3) is used to rank the documents for a given query. The probability of words can be estimated using the maximum-likelihood criterion as

$$\begin{aligned} p_{ml}(q_i|\hat{\theta}_Q) =& \frac{c_i(\mathbf{q})}{M_q} \\ p_{ml}(d_i|\hat{\theta}_D) =& \frac{c_i(\mathbf{d})}{M_d}, \end{aligned} \tag{5.5}$$

where $c_i(\mathbf{q})$ and $c_i(\mathbf{d})$ are the term frequencies of the words q_i and d_i in the query and a document, respectively.

In the empirical estimation of the document models as specified above, the probability of the visual words that do not occur in a document will be zero. This

will lead to infinite numbers in the relevance estimation based on KL divergence in Eqn. (5.3). Hence, smoothing [138][31] should be introduced to address this problem, similarly as in the case of speech recognition and machine translation.

As suggested in the study of language modeling approaches for image retrieval [31], the Jelinek-Mercer smoothing method performs the best in the image retrieval context and we therefore adopt it in this chapter. The Jelinek-Mercer smoothing is defined as a linear interpolation of the maximum likelihood estimation of the language model and the collection model, which can be formulated as

$$p_\lambda(w_i|\hat{\theta}_D) = (1-\lambda)p_{ml}(w_i|\hat{\theta}_D) + \lambda p(w_i|\mathcal{C}), \tag{5.6}$$

where $p(w_i|\mathcal{C})$ is the collection language model and $\lambda \in [0,1]$ is the trade-off parameter to control the contribution of the smoothing term.

5.3.3 Contextual object retrieval model

Standard works on object retrieval based on the bag-of-visual words image representation [77][31] use the visual words located within the bounding box to estimate the query model. However, as stated above, the visual context can be used to improve the reliability of this estimation by looking beyond the bounding box information only. In that case, the KL divergence based retrieval model introduced in Section 5.3.2 can again be directly applied to estimate the relevance between the query and database images, but now employing a better, context-aware query model.

In our COR model we assume that the the query image with its bounding box specification is generated from the following distribution:

$$\begin{aligned} p(\mathbf{q}|\theta_Q) =& p(\mathbf{q}^I, \mathbf{q}^b|\theta_Q) \\ \propto& \prod_{i=1}^{M_q} p(q_i, p_i|\theta_Q), \end{aligned} \tag{5.7}$$

with

$$p(q_i, p_i|\theta_Q) = p(q_i|\theta_Q)^{S(p_i,\mathbf{q})}, \tag{5.8}$$

where $S(p_i, \mathbf{q})$ is the search intent score of the visual word q_i at the position p_i. While the COR model unifies both the visual words from the bounding box and the visual words from the context of the bounding box for inferring a more reliable query model, the search intent score steers this inference process by indicating the confidence of a given visual word to be relevant to the search intent. As a comparison, for the language modeling approach without considering the context, the query also follows the above distribution, however with a binary search intent score. For the visual words located within the bounding box, this score is then equal to 1, while being 0 otherwise.

Based on the distribution (5.7) the maximum likelihood estimation of the context-aware query model θ_Q can then be derived as

$$p(w_j|\theta_Q) = \frac{\sum_{i=1}^{M_q} S(p_i, \mathbf{q})\delta(q_i = w_j)}{\sum_{i=1}^{M_q} S(p_i, \mathbf{q})}, \tag{5.9}$$

and then integrated into the retrieval model (5.3) to rank the images.

In the following section we will introduce two methods to estimate the search intent score $S(p_i, \mathbf{q})$.

5.4 Search intent score estimation

The search intent score of each visual word in the query image can be defined to be proportional to the probability of the corresponding position of that visual word to reflect the user's search intent given the query image and the bounding box:

$$S(p_i, \mathbf{q}) \propto p(p_i|\mathbf{q}). \tag{5.10}$$

Based on Bayes' formula and assuming a uniform prior, the probability (5.10) is proportional to the likelihood of generating the query image and the bounding box given the search intent score:

$$\begin{aligned} p(p_i|\mathbf{q}) =& p(p_i|\mathbf{q}^I, \mathbf{q}^b) \\ \propto& p(\mathbf{q}^I, \mathbf{q}^b|p_i). \end{aligned} \tag{5.11}$$

We realistically assume that the bounding box and the query image are conditionally independent given the search intent score per position. Then, we can write

$$p(p_i|\mathbf{q}) \propto p(\mathbf{q}^b|p_i)p(\mathbf{q}^I|p_i). \tag{5.12}$$

Applying the Bayes' formula, we can transform the previous expression as

$$p(p_i|\mathbf{q}) \propto p(p_i|\mathbf{q}^b)p(p_i|\mathbf{q}^I). \tag{5.13}$$

The first term, $p(p_i|\mathbf{q}^b)$, is the probability of the position p_i to reflect the search intent as inferred from the bounding box. The second term, $p(p_i|\mathbf{q}^I)$, can be seen as the probability of the position p_i to represent salient properties of the query image in general, thus not related to any specific search session, but indicating a logical choice of a prior for inferring a user's search intent given a search session. As such, the second term in Eqn. (5.13) can be estimated through saliency detection and used to improve the reliability of the search intent score estimation, especially when the bounding box specification is unreliable.

In the next steps, we will estimate the prior using the saliency detection method [64] in Section 5.4.1 and then propose two algorithms to compute the search intent score from the bounding box in Section 5.4.2.

5.4.1 Saliency detection

While human perceptual attention analysis plays an important role in the biological vision and human cognition, computational attention analysis or saliency detection is an important approach in content-based image retrieval to detecting the potentially important and representative regions in an image [64]. As such, saliency detection can be regarded as a means for determining the prior for defining the ROI in a given query image. Among the existing methods for computing the saliency of an image [105][39][64], we adopted the contrast-based image attention analysis [64] due to its simplicity and computational efficiency. According to this approach, color contrast plays a critical role in attracting human attention when looking at an image. We compute a contrast-based saliency score for each of the positions in an image using the following expression:

$$C_i = \sum_{y \in \mathcal{N}_i} d(\mathbf{l}(p_i), \mathbf{l}(y)), \tag{5.14}$$

where $\mathcal{N}_i$ is the neighborhood of the position p_i in the image, and $\mathbf{l}(p_i)$ and $\mathbf{l}(y)$ are the color values in the positions p_i and y. According to the suggestions from [64], we worked with the colors in the LUV space. Finally, d is the Gaussian distance between the color values. Normalization of the color contrast C_i into the range $[0, 1]$ leads to the saliency score A_i for each of the positions in an image. We then transform the saliency score into the prior probability based on the Gibbs distribution using the expression

$$p(p_i|\mathbf{q}^I) \propto \exp(-\gamma(A_i - 1)^2), \tag{5.15}$$

where γ is the inverse of temperature.

5.4.2 Search intent from the bounding box

Search intent score estimation from the bounding box is an important ingredient of the proposed contextual object retrieval model. In the following, we will present two approaches to estimating the search intent, one from the spatial propagation (dual-sigmoid approximation) and the other from the appearance propagation perspective (appearance-propagation based method using matting)

Search intent from the bounding box by dual-sigmoid approximation

The bounding box specification is an uncertain event conditioned by the unobserved search intent. Users normally intend to specify a rough and inaccurate bounding box to save the effort. Due to its limiting rectangular shape the bounding box is unlikely to accurately represent a complex ROI.

In our approach to estimating the search intent score based on the information from the bounding box we first assume that the intents for the two dimensions of

the image are independent of each other so that the intent probability can be decomposed into the product of the probabilities estimated from the two dimensions respectively:

$$\begin{aligned} p(p_i|\mathbf{q}^b) =& p(x_i, y_i|x_l, y_l, x_r, y_r) \\ =& f(x_i; x_l, x_r, \delta) f(y_i; y_l, y_r, \delta). \end{aligned} \tag{5.16}$$

The function f, i.e., the search intent score estimation from a single dimension, should be a smoothed approximation of the bounding box along that dimension in order to take the uncertainty and the context into consideration. That is, the value of f for $x_l < x_i < x_r$ should be close to 1 and should be approaching 0 the further x_i is from the bounding box. To obtain a probability distribution, we propose to model f as the minimization of two sigmoid functions for the two sides of the bounding box along each dimension. For the x-dimension this model can be defined as

$$\begin{aligned} f(x_i; x_l, x_r, \delta) = \min \big(& \frac{1}{1 + \exp(\delta(x_l - x_i))}, \\ & \frac{1}{1 + \exp(\delta(x_i - x_r))} \big), \end{aligned} \tag{5.17}$$

where δ is a parameter serving as a tradeoff between fitting the bounding box and being sufficiently smooth to incorporate the context. We use the same model for f in the y-dimension as well.

Figure 5.4 illustrates the dual sigmoid function defined in Eqn. (5.17). We can see that the smooth dual sigmoid function indeed approximates the bounding box. The approximation is better with a larger δ. In particular when $\delta \to +\infty$ the function f becomes equal to the bounding box specification. Smaller δ leads to more smoothing, which means that the bounding box specification is more uncertain and the context information has a larger effect. In the extreme case of $\delta = 0$ the bounding box specification is unused and we just use the whole image as the query.

Finally, we define the search intent score to be the product of the prior (5.15) and the probability estimation indicating the search intent based on the bounding box specification:

$$\begin{aligned} S_a(p_i, \mathbf{q}) \stackrel{\text{def}}{=\!=} & \exp(-\gamma(A_i - 1)^2) \times \\ & f(x_i; x_l, x_r, \delta) f(y_i; y_l, y_r, \delta). \end{aligned} \tag{5.18}$$

The parameters γ and δ control the contributions from the prior and the bounding box to the intent score estimation. If the bounding box specification is reliable then we should adopt a smaller γ and a larger δ, and vice versa.

Search intent from the bounding box by matting

The goal of estimating the search intent score from the bounding box is to assign high scores to the object of interest, which is normally in the foreground. Low

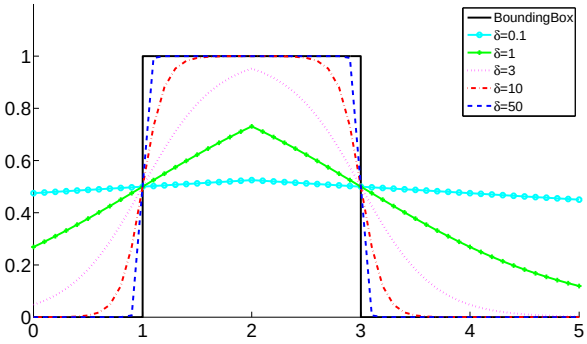

Figure 5.4: *Illustration of the dual sigmoid function for the search intent score estimation based on the bounding box.*

scores should be assigned to the background or other foreground objects of no interest for the user, which, however, should still be regarded as a useful context for the image retrieval based on the ROI. This is similar to the problem of image matting, which is to separate the foreground from the background by estimating alpha values (the values in the alpha channel to indicate the opacity) for each pixel [112]. Hence, the image matting algorithms can be adapted directly to estimate the search intent from the bounding box.

Since the bounding box is a rough specification of the object of interest, it can be regarded as containing the object of interest together with some irrelevant image parts, which are—due to the immediate adjacency to the target object—most likely to contain the background and not some other parts of the foreground object. We now adopt a three-step approach to determine which parts of the bounding box belong to the foreground and which to the background. Firstly, we segment the image guided by the bounding box specification and estimate the foreground and background models. Then we employ the estimated models to select the pixels which most probably belong to either the foreground or to the background. Finally the alpha value or the search intent score of each pixel is estimated based on the pseudo-foreground and pseudo-background pixels.

The segmentation in the first step is performed using the GrabCut algorithm, where the foreground and background models are Gaussian Mixture Models (GMM) in the RGB color space [83]. The objective of GrabCut is to minimize an energy function comprising a data fitting term and a smoothness term, as defined by the following expression:

$$E(\alpha, \mathbf{k}, \theta, \mathbf{z}) = U(\alpha, \mathbf{k}, \theta, \mathbf{z}) + \gamma V(\alpha, \mathbf{z}), \tag{5.19}$$

Here, $\alpha \in \{\mathcal{F}, \mathcal{B}\}$ indicates whether the pixels belong to the foreground or background, and $\mathbf{k}$ indicates which GMM is assigned. Furthermore, θ are the model

parameters, $\mathbf{z}$ represents the color of each pixel, and γ is the parameter regulating the trade-off between data fitting and smoothing.

Specifically, the data fitting term U and the smoothness term V are defined respectively as follows:

$$\begin{aligned} U(\alpha, \mathbf{k}, \theta, \mathbf{z}) = \sum_n -\log \pi(\alpha_n, k_n) + \frac{1}{2}\log|\Sigma(\alpha_n, k_n)| + \\ \frac{1}{2}[z_n - \mu(\alpha_n, k_n)]^T \Sigma(\alpha_n, k_n)^{-1}[z_n - \mu(\alpha_n, k_n)], \end{aligned} \tag{5.20}$$

$$V(\alpha, \mathbf{z}) = \sum_{(m,n)\in\mathbf{C}} [\alpha_m \neq \alpha_n] \exp(-\beta||z_m - z_n||^2) \tag{5.21}$$

where μ and Σ are the parameters of the foreground and background models, where π is the mixture weight, where $\mathbf{C}$ is the set of pairs of neighboring pixels, and where β is the parameter to adjust the extent of smoothness in a coherent region.

After GrabCut, which is based on an iterative energy minimization algorithm [83], is completed, we use the estimated foreground and background models to obtain the probabilities of each pixel x belonging to the foreground and background. Such probability for the foreground is defined as

$$P_{\mathcal{F}}(x) = \frac{P(x|\theta, \mathcal{F})}{P(x|\theta, \mathcal{F}) + P(x|\theta, \mathcal{B})}. \tag{5.22}$$

Directly using the estimated probabilities as the search intent scores does not take into account the spatial smoothness and may not perform well. We therefore compute the intent scores based on selected pseudo-foreground and pseudo-background pixels. Specifically, we select the top 10% pixels inside the bounding box having the largest foreground probabilities and the 20% pixels outside the bounding box having the largest background probabilities as the pseudo-foreground $\Omega_{\mathcal{F}}$ and pseudo-background pixels $\Omega_{\mathcal{B}}$, respectively, which serve as input to the matting algorithm.

The matting algorithm adopted in this chapter is based on the geodesic distance [10], as defined by the following expression:

$$D_l(x) = \min_{s\in\Omega_l} d(s, x), \tag{5.23}$$

where $l \in \{\mathcal{F}, \mathcal{B}\}$ and where $d(s, x)$ is computed as follows:

$$d(s, x) = \min_{P_{s_1,s_2}} \int_0^1 |W \cdot P_{s_1,s_2}(p)|dp, \tag{5.24}$$

where P_{s_1,s_2} is any path connecting the two pixels s_1 and s_2; $W = \nabla P_{\mathcal{F}}(x)$.

In view of the above, the search intent score from the bounding box that incorporates matting and the prior can be computed as

$$S_m(p_i, \mathbf{q}) \stackrel{\text{def}}{=\!=} \exp(-\gamma(A_i - 1)^2) \times \frac{D_{\mathcal{B}}(x_i)}{D_{\mathcal{F}}(x_i) + D_{\mathcal{B}}(x_i)}, \tag{5.25}$$

where γ controls the contribution from the prior and x_i is the pixel value in the poisition p_i.

By integrating the query language model in Eqn. (5.9) with the search intent score estimated by Eqn. (5.18) and Eqn. (5.25) into the KL divergence retrieval model (5.3), we obtain two contextual object retrieval models referred to as COR_a and COR_m that will be evaluated experimentally in the next section.

5.5 Experiments

5.5.1 Datasets

To evaluate the proposed COR models, we performed experiments on three representative image datasets: **Oxford5K**, **Oxford5K+ImageNet500K** and **Web1M**.

The **Oxford5K dataset** [77][6] was collected from Flickr [4] by using 17 Oxford landmarks as queries. In total, 5062 images have been acquired, among which 55 images comprising 11 landmarks were selected as query images. The bounding box specifying the ROI was inserted manually on the query images. Furthermore, the entire dataset was annotated based on the relevance with respect to the 55 query images, which provided the ground truth for our experimental evaluation. The SIFT features for all images were extracted using the publicly available software tool[2] and then quantized using the 1M visual vocabulary to visual word representations.

To evaluate the proposed method in a large-scale image retrieval setting, we constructed the **Oxford5K+ImageNet500K** dataset by combining the Oxford5K and a part of the ImageNet dataset [27]. Specifically, we sampled 500K images from about 10M images in ImageNet, and then combined them with the 5K images in Oxford5K into a new collection. We still used the 55 query images and the associated bounding boxes in Oxford5K as queries. We assumed that the 500K ImageNet data contain no relevant images to the 55 queries. The SIFT features for the newly added 500K images were extracted and quantized into visual words using the same procedure as for Oxford5K.

The query images in the Oxford5K and Oxford5K+ImageNet500K datasets are all about the landmarks at Oxford University, which can be considered only a narrow image retrieval use case. To evaluate the performance of our COR models on a large-scale dataset able to match more versatile search intents, we created the **Web1M** image collection comprising 1 million of mostly clicked images on

[2] http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings

the Web. 12 keywords[3] covering different categories, i.e., animals, landmarks, paintings, logos, and book covers, were used to collect near-duplicate images from Flickr, which are then added to the database together with the aforementioned 1M popular images. From all these images, 45 images were randomly selected as query images and then manually marked by bounding boxes to indicate possible objects of interest. This resulted in 56 example objects that were used as queries to evaluate the proposed models. Again the visual word representation was obtained here in the same way as in the other two datasets.

5.5.2 Experimental setup

In order to demonstrate the effectiveness of the proposed COR models, we compare them with two baseline object retrieval models. One is the cosine model (Cosine), which is based on the cosine similarity between the vector models of the query and the database images. The other is the general (context-unaware) language modeling approach (LM) that we introduced in Section 5.3.2. Each of the baseline methods only uses the visual words inside the bounding box for building the query model, while the contextual visual words are ignored.

In addition to a comparative analysis involving the methods listed above, we also included in the evaluation two variants of the matting-based COR model. The first variant is the COR_g model where the foreground probability is computed using the GMM models estimated by GrabCut as defined in Eqn. (5.22):

$$S_g(p_i, \mathbf{q}) \stackrel{\text{def}}{=\!=} \exp(-\gamma(A_i - 1)^2) \times P_{\mathcal{F}}(x). \tag{5.26}$$

The second variant is the COR_w model that uses the alpha values computed based on the weighted foreground probability proposed in [10]:

$$\begin{aligned} &S_w(p_i, \mathbf{q}) \stackrel{\text{def}}{=\!=} \exp(-\gamma(A_i - 1)^2) \times \frac{\omega_{\mathcal{F}}(x_i)}{\omega_{\mathcal{F}}(x_i) + \omega_{\mathcal{B}}(x_i)}, \\ &\omega_l(x_i) = D_l(x_i)^{-1} \cdot P_l(x_i),\ l \in \{\mathcal{F}, \mathcal{B}\}. \end{aligned} \tag{5.27}$$

The parameters of the models, such as the λ for Jelinek-Mercer smoothing, γ for saliency weight re-scaling and δ for bounding box weight re-scaling, were selected to optimize the average performance over all queries. The effects of γ and δ on the retrieval performance will be analyzed in more details in Section 5.5.5.

The models were evaluated in terms of the Average Precision (AP), which is defined as the average of the precision values computed at various recall levels. The AP over all queries were then averaged to obtain the Mean Average Precision (MAP).

[3]The 12 keywords include *Big Ben, Eiffel Tower, Leaning Tower of Pisa, Ferrari, Starbucks, Uncle Sam, Mocking jay, Leopard, Panda, Zebra, Mona Lisa, Starry Night*

Table 5.1: *The MAP of the six methods on Oxford5K, Oxford5K+ImageNet500K, and Web1M datasets.*

Dataset	Cosine	LM	$\mathrm{COR_g}$	$\mathrm{COR_w}$	$\mathrm{COR_m}$	$\mathrm{COR_a}$
Oxford5K	0.614	0.623	0.581	0.591	0.611	**0.659**
Oxford5K+ImageNet500K	0.418	0.581	0.546	0.552	0.578	**0.621**
Web1M	0.414	0.644	0.587	0.601	0.651	**0.700**

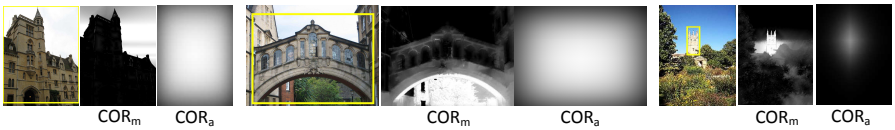

Figure 5.5: *Illustration of three cases where the matting-based retrieval models will not estimate a good intent score. For each case, the left image is the query example with the bounding box indicated by the yellow rectangle. The right two images show the intent score maps computed using Eqn. (5.18) and Eqn. (5.25) ($\gamma = 0$) for* $\mathrm{COR_a}$ *and* $\mathrm{COR_m}$*, respectively. The brighter the point's intensity, the large is the search intent score, and vice versa.*

5.5.3 Performance comparison on two Oxford landmark datasets

Comparison of different COR models

The MAP values obtained for all models on the two Oxford landmark datasets are shown in Table 5.1. It can be observed that the three matting based methods all perform worse than the LM baseline, which states that the search intent scores computed using matting do not reflect the user's search intent behind the query specification. Taking a look at the query images leads to the following possible explanation of this result. First, the users' bounding box specification may be inaccurate, so that it includes some of the background regions, such as the cloud in Fig. 5.5 (a). Then the estimated foreground model may get confused by the background, which leads to inaccurate intent score estimation. Second, since the matting methods are based on color coherence, the foreground may receive low intent scores when the background has similar color appearance as the foreground, as shown in Fig. 5.5 (b). Due to the same reason, background may get assigned incorrectly high intent scores, as shown in Fig. 5.5 (c).

Among the three matting based approaches including $\mathrm{COR_g}$, $\mathrm{COR_w}$, and $\mathrm{COR_m}$, $\mathrm{COR_m}$ significantly outperforms the others. This indicates that the proposed three-step matting-based retrieval model is more effective than the other two more simplistic solutions and that the geodesic distance based search intent score estimation is better than the one based on the foreground probability. We

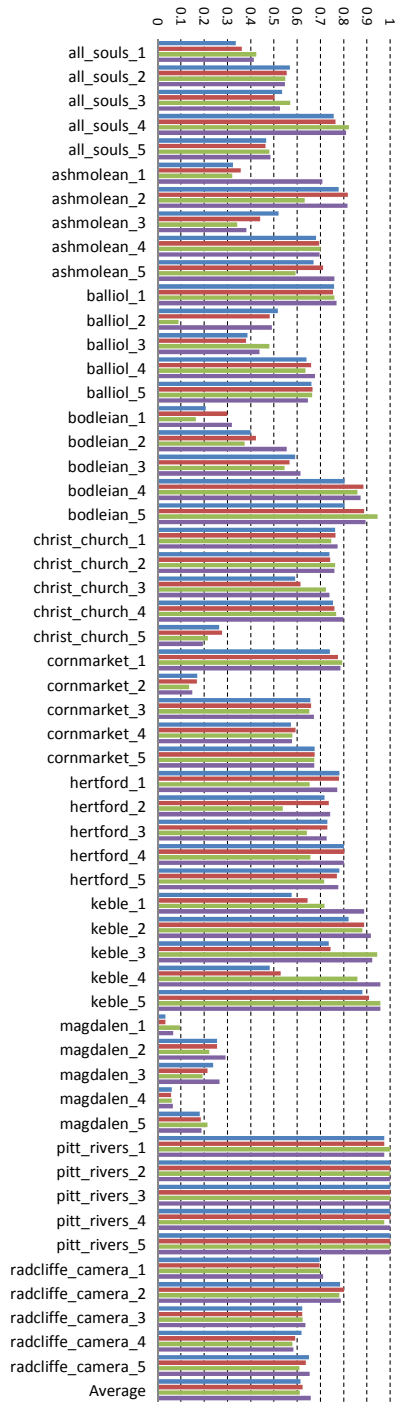

Figure 5.6: *The AP per query image on Oxford5K dataset.*

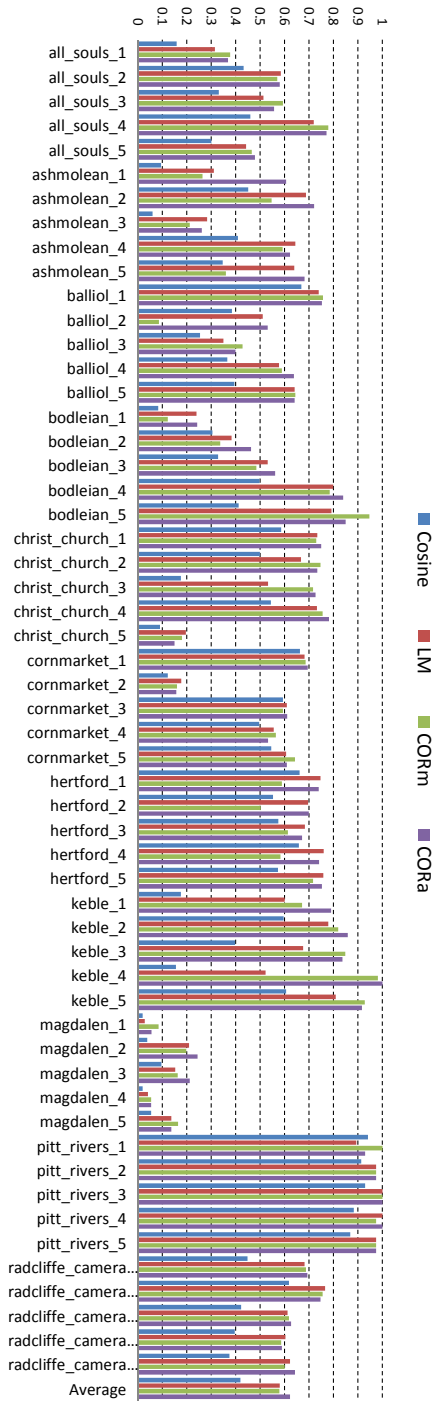

Figure 5.7: *The AP per query image on Oxford5K+ImageNet500K dataset.*

argue that the main reason for this result is that it is difficult to have a good estimation of the foreground and background model when the background is complex or when the bounding box is very inaccurate.

In the remaining evaluation steps, we will use COR_a as the representative of the contextual object retrieval models to compare with context-unaware baselines and in this way gain insights about the usefulness of visual context for object retrieval as introduced in this chapter.

Comparison with non-contextual methods

Observing the results obtained for Oxford5K as shown in Table 5.1, we can conclude that the language modeling approach is slightly better than the cosine model. This is because the language modeling approach has more solid theoretical foundations, as have already been demonstrated in the field of information retrieval. The proposed COR_a model significantly boosts the performance, with 7.1% and 5.5% relative improvement over the cosine and language models, respectively. On the Oxford5K+ImageNet500K dataset the improvement of COR_a over Cosine and LM is 21.8% and 6.9% respectively. These results suggest that the visual context surrounding the ROI in the query image is indeed a useful auxiliary source of information, which may lead to a more reliable relevance estimation between the query and database images. The results also indicate that the proposed COR model successfully leverages the context information into the retrieval model. Performance comparisons for each of the 55 query images on Oxford5K and Oxford5K+ImageNet500K datasets are shown in Fig. 5.6 and Fig. 5.7, respectively.

The usefulness of visual context

The performance for each of the 11 landmarks on the Oxford5K dataset are shown in Fig. 5.8. We can observe an improvement of COR_a over LM for 7 out of 11 landmarks. Among them "keble" improves most significantly, by 32.0% and 25.6% over the cosine and LM model, followed by "magdalen" (16.0% and 19.8%) and "ashmolean" (13.8% and 12.5%). The result on Oxford5K+ImageNet500K dataset as shown in Fig. 5.9 suggests a similar conclusion. When we take a look at the query images and bounding boxes of such queries, all of which can be found in [6], we find that the bounding boxes for the queries with the most performance improvements tend to be small. Since small bounding boxes often contain few visual words, the relevance estimation based on these few features in the query region is likely to be unreliable. This supports our hypothesis that the visual context information can improve the reliability of relevance estimation in these difficult cases and in this way improve the retrieval performance.

For the other 4 landmarks, i.e., "pitt_rivers", "hertford", "radcliffe_camera" and "cornmarket", the introduction of visual context did not improve the retrieval performance. The reason is that for these 4 landmark queries the bounding boxes

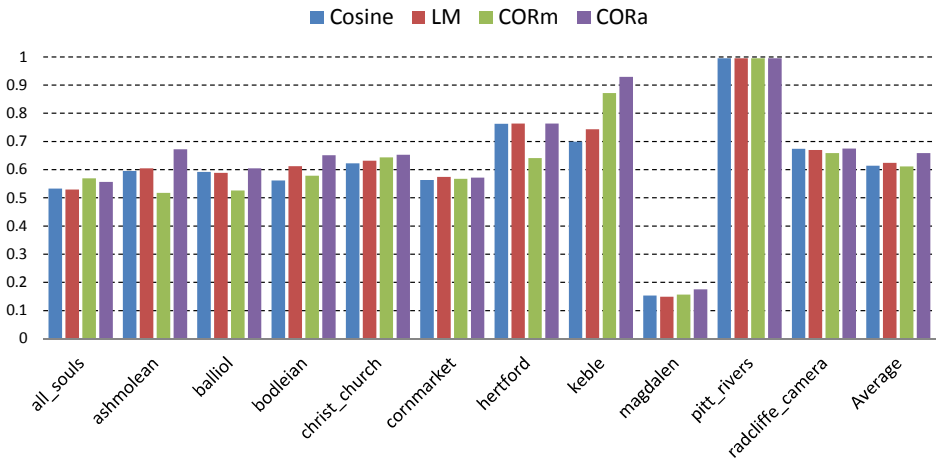

Figure 5.8: *The AP for different landmarks on Oxford5K dataset.*

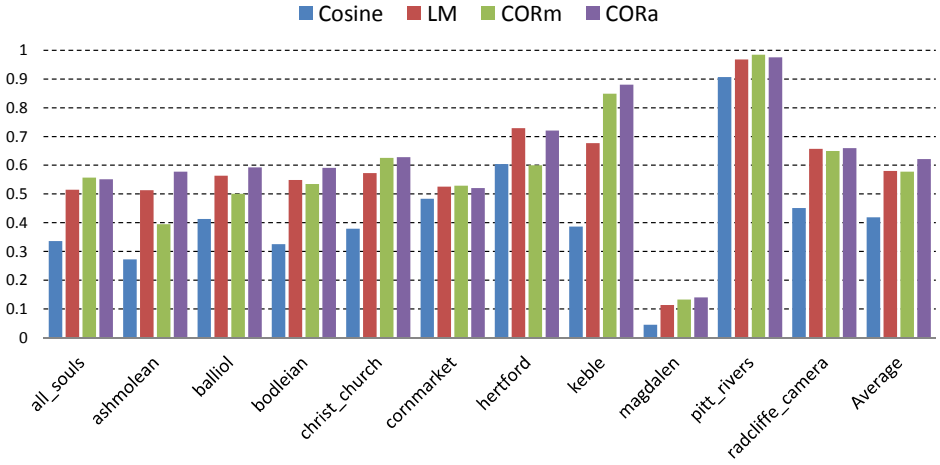

Figure 5.9: *The AP for different landmarks on Oxford5K+ImageNet500K dataset.*

are larger than in other cases and therefore more informative. Consequently, the context did not play a large role there. While the context is less useful in such cases, the COR_a model still achieves a comparable performance with the LM method. The largest performance degradation of COR_a compared with LM is only 0.39% ("cornmarket"). One of the reasons for the small degradation is that we use the same values for the parameters γ and δ. While the values are optimized for the average performance it will not be optimal for some of the queries. A more elaborate analysis of the dependence of the COR_a performance variation on a varying bounding box size can be found in Fig. 5.10.

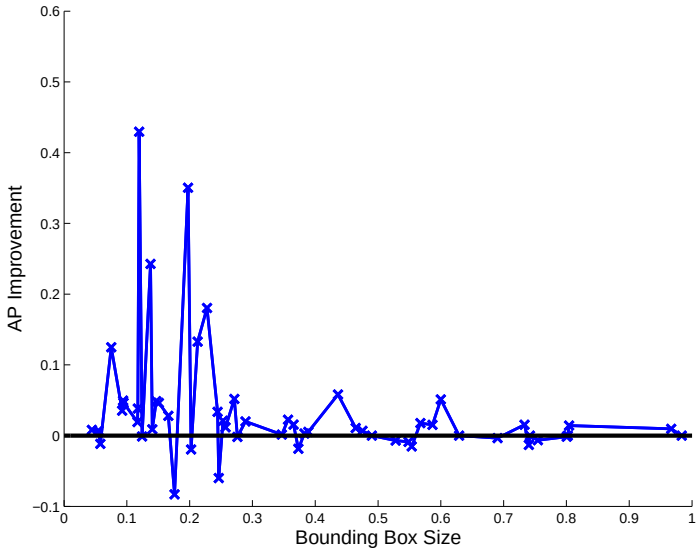

Figure 5.10: *Performance improvement of* $\mathrm{COR_a}$ *model varies with varying bounding box sizes on Oxford5K. Absolute MAP improvement of* $\mathrm{COR_a}$ *over LM can be observed.*

Finally, we show in Fig. 5.11 the retrieved images for three sample queries on Oxford5K. We can see that although the three queries clearly benefit from the visual context, they draw this benefit in different ways. For "keble_4" and "magdalen_2", where the query formulation from the visual words within the small bounding box is uncertain, $\mathrm{COR_a}$ manages to remove the completely irrelevant images (marked with blue rectangles) from the top results and improves the retrieval performance. For the query "pitt_reivers_4", even if the introduction of context does not boost the MAP, it still can replace the completely irrelevant image (the image of people marked with blue rectangle) with the other image which is more consistent to the query (the 9th image in the result for "pitt_reivers_4", which is a building, just like the query). This demonstrates the power of visual context in improving the reliability of relevance estimation.

5.5.4 Performance comparison on Web1M dataset

Evaluation of the retrieval performance on Web1M presented in Table 5.1 again leads to the conclusion that $\mathrm{COR_a}$ performs significantly better than the context-unaware models. It achieves 69% relative improvement over the cosine model and 8.7% relative improvement over LM, which demonstrates the usefulness of the visual context in a broad range of image retrieval use cases. Among the four contextual models, $\mathrm{COR_a}$ obviously outperforms the others and certainly can be regarded as the most effective way to incorporate visual context information into

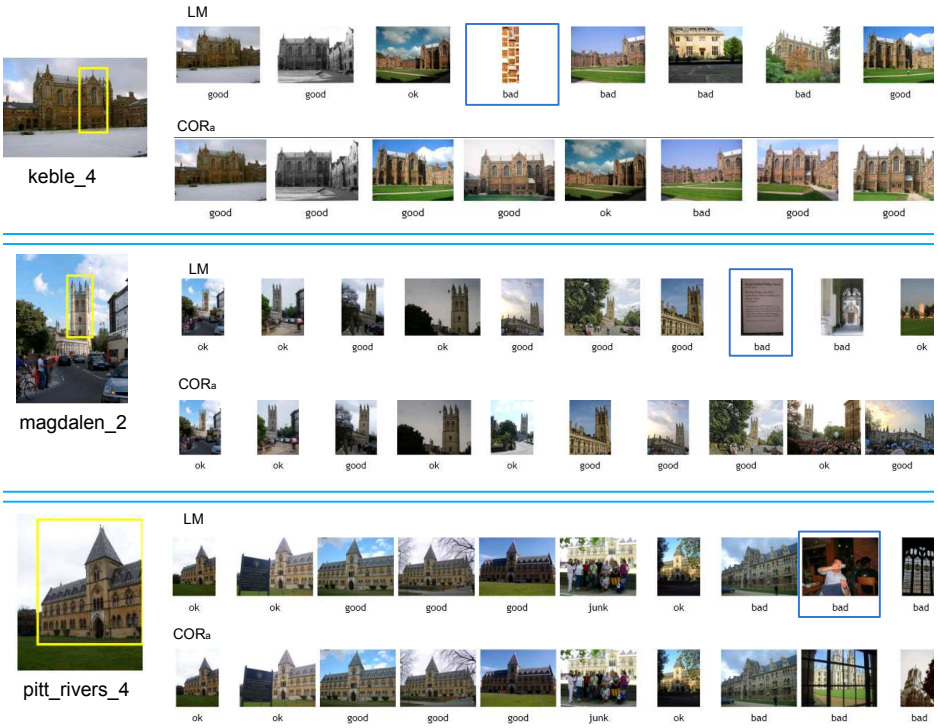

Figure 5.11: *Sample search results for* $\mathrm{COR_a}$ *and LM on Oxford5K dataset. The yellow rectangle on the query image is the bounding box to specify the object of interest.*

object retrieval.

Another noteworthy observation is that on Web1M $\mathrm{COR_m}$ performs slightly better than the LM baseline. By analyzing the average performance of each query category in Fig. 5.12, we can conclude that the overall performance of $\mathrm{COR_m}$ is boosted mostly because it has a large performance gain over LM on the query category "Big Ben". Since some of the bounding boxes for "Big Ben" are on the clock instead of the entire building, the color-coherence based matting works well to propagate the intent to the remaining parts of the building and in this way improves the retrieval performance. Fig. 5.13 shows the sample results to illustrate this observation. The AP per query object is shown in Fig. 5.14.

The results of our study of the variance in the usefulness of visual context on various search intent categories including landmark, animal, logo, book cover, and painting are summarized in Table 5.2. We can see that the largest performance improvement by incorporating visual context is achieved on landmarks. We argue that this may be because landmarks are usually in a fixed geographical location. Therefore their visual context, such as the adjacent landmarks, is relatively more stable and better capable to improve the retrieval performance.

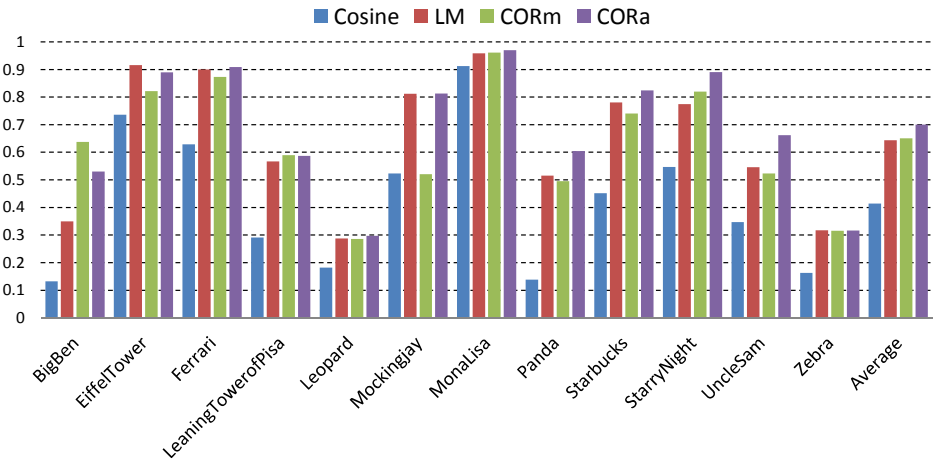

Figure 5.12: *The AP for different queries on Web1M dataset.*

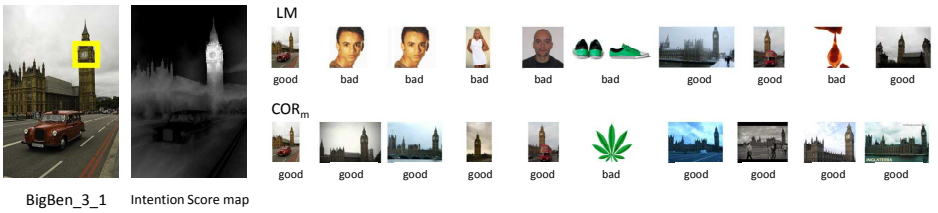

Figure 5.13: *The example query "BigBen_3_1" for which* $\mathrm{COR_m}$ *achieves significant performance improvement.*

Table 5.2: *The MAP of each search intent category on Web1M dataset.*

Category	LM	$\mathrm{COR_a}$	
		MAP	Gain
Landmark	0.546	0.634	16.21%
Animal	0.352	0.379	7.59%
Logo	0.767	0.815	6.34%
Painting	0.866	0.930	7.35%
Book Cover	0.812	0.813	0.17%

Although performing not as well as landmarks, queries on animals, logos, and paintings still show a significant performance boost when visual context is taken into account. The relatedness of these objects to the context, though not as strong

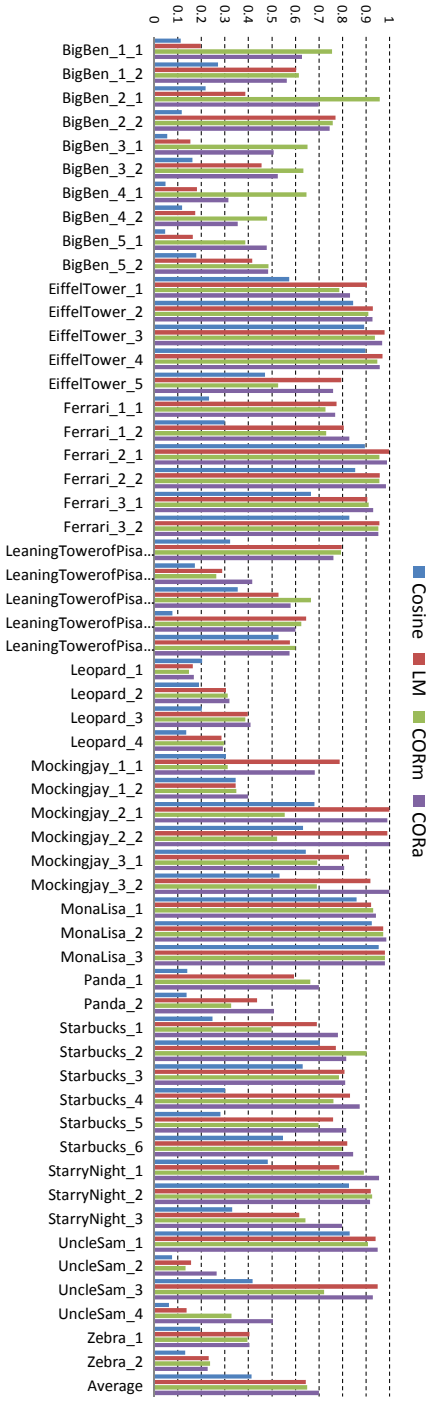

Figure 5.14: *The AP per query image on Web1M dataset.*

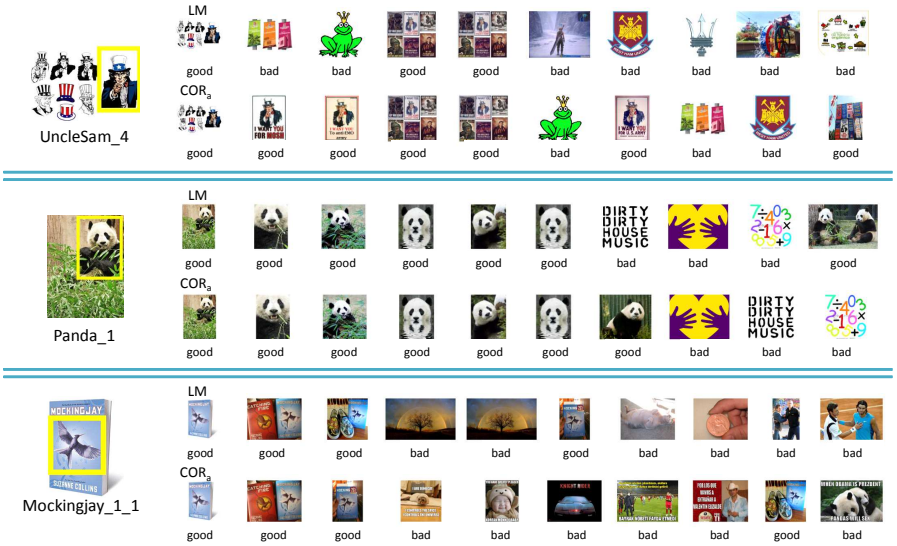

Figure 5.15: *The sample search results for* $\mathrm{COR_a}$ *and LM on Web1M dataset. The yellow rectangle on the query image is the bounding box to specify the object of interest.*

as landmarks, can be regarded as helpful information to disambiguate the search intent. For example, the panda often co-occurs with bamboos and the paintings are often fixed in a frame. Fig. 5.15 illustrates this and shows some results on these types of queries.

For the book-cover queries, the contextual object retrieval model shows comparable performance with LM. The reason is that in book-cover images, there are usually texts around the query object, such as the "Mockingjay_1_1" shown in Fig. 5.15. Since the text contains strong visual patterns resulting in a large quantity of SIFT points, these contextual features on the text may become too dominant and disturb the retrieval performance. This search intent category represents the cases in which the context shows weak correlation to the search intent of object retrieval and is hardly useful to further improve the retrieval performance. However, the introduction of visual context doesn't deteriorate the performance, too. Since the search intent scores tend to be small for the contextual words, the context cannot play a significant role if it is weakly correlated with the object of interest.

5.5.5 Parameter analysis

We conclude the experimental section by analyzing in more detail the effects of the two parameters in our proposed $\mathrm{COR_a}$ model that steer the search intent score computation on the object retrieval performance. These parameters are γ

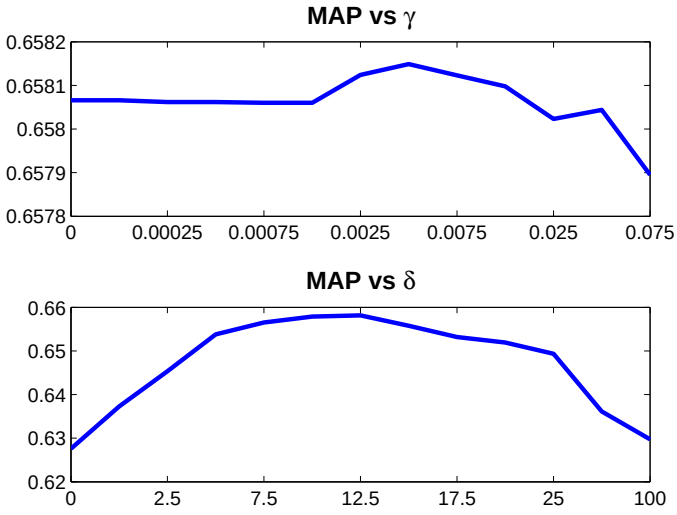

Figure 5.16: *The effect of the parameters γ and δ on the retrieval performance.*

and δ. Larger γ means that the retrieval result is more sensitive to the detected saliency, while smaller γ suggests that saliency plays a smaller role in the retrieval model. The parameter δ models the reliability of the bounding box specification. Larger δ means that we rely more on the user-specified bounding box. The case $\delta \to +\infty$ makes the $\mathrm{COR_a}$ model identical to LM, while $\delta = 0$ means that the whole image is used as a query, without considering the bounding box to restrict the query region.

The results on Oxford5K dataset as shown in Fig. 5.16 indicate the optimal values for γ and δ. Setting γ to about 0.005 and δ to around 10 achieves the best performance. Since in this dataset the bounding boxes are annotated for research purposes, they are expected to be much more accurate than the bounding box specified by an average user performing a general search task. The more accurate the bounding box is, the larger the optimal δ will be. Then, the saliency becomes less important since the prior information plays only a small role when the bounding box is accurate. In a practical system, the system administrator can adjust the parameters based on the user behavior.

5.6 Conclusions and Future Work

In this chapter we proposed two contextual object retrieval models to improve the object retrieval performance when the query object is specified by a rectangular bounding box in the query image. Since the bounding box is an uncertain obser-

vation of the search intent, we infer this intent for each of the visual words in the query image from the bounding box specification and from the prior estimated from the query image's saliency map. Then the search intent scores are integrated into the COR model to more effectively meet users' true information needs. Experiments on several datasets demonstrate that one of the proposed COR models, the one based on spatial propagation by using dual-sigmoid approximation, is particularly effective in improving the object retrieval performance.

The idea of the contextual object retrieval model introduced in this chapter can be regarded as a preliminary work on a more general topic: context-aware Multimedia Information Retrieval (MIR) [63]. While content-based [23][95] and concept-based [96] MIR have made a great progress in the past years, context-aware MIR can be seen as a new paradigm to jointly address the semantic gap [23][95][96] and intention gap [136] challenges in MIR by incorporating the contextual information. In addition to the contextual information used in this paper, various other kinds of context, e.g., the text surrounding the query image in the Web page and the user click-through log from a search engine, will be investigated to evaluate their role in MIR in the future.

Chapter 6

Video-based Image Retrieval

[1]

Despite the possibility to take the visual context of the query object into account, as proposed in the previous chapter, the performance of image retrieval solutions based on the query-by-example (QBE) principle may still vary significantly due to the likely variations in the capture conditions (e.g. light, blur, scale, occlusion) and viewpoint between the query image and the images in the collection. In this chapter, we propose a framework in which some of these variations are explicitly addressed to improve the reliability of QBE-based image retrieval. We aim at the use scenario involving the user capturing the query object by his/her mobile device and requesting information augmenting the query from the database. Reliability improvement is achieved by allowing the user to submit not a single image but a short video clip as a query. Since a video clip may combine object or scene appearances captured from different viewpoints and under different conditions, the rich information contained therein can be exploited to discover the proper query representation and to improve the relevance of the retrieved results. The experimental results show that video-based image retrieval (VBIR) is significantly more reliable than the retrieval using a single image as query. Furthermore, to make the proposed framework deployable in a practical mobile image retrieval system, where real-time query response is required, we also propose the priority queue-based feature description scheme and cache-based bi-quantization algorithm for an efficient parallel implementation of the VBIR concept.

[1]This chapter was published as: Linjun Yang, Yang Cai, Alan Hanjalic, Xian-Sheng Hua, Shipeng Li, "Searching for images by video." International Journal of Multimedia Information Retrieval, 2012 [122].

6.1 Introduction

Image retrieval based on the Query-by-Example (QBE) principle has recently been revived and gained increasing attention from both the research community and industry. A probable reason lies in the success of the applications like Google Goggles[2], TinEye[3], and "Find more sizes" of Bing Image Search[4] that have become popular tools for retrieving images or other information related to the visual example serving as query. In particular, in the mobile use scenario, where the user can easily capture the visual query using the camera on his/her mobile device, QBE-based image retrieval appears as a highly convenient retrieval concept, as opposed to the one requiring textual keywords as queries.

Despite extensive research efforts in the past, QBE-based image retrieval is still insufficiently reliable, largely because of the likely variations in the capture conditions (e.g. light, blur, scale, occlusion) and viewpoint between the query image and the images in the collection. This query-collection mismatch has been difficult to resolve due to the still imperfect visual features used to represent the query and the collection images. While, for instance, the SIFT features [61] are effective in general, they are still insufficiently capable of handling the variations in blur and occlusion. Furthermore, in a typical SIFT-based image representation using visual words [93], the visual word quantization degrades the retrieval reliability to trade off for the scalability of the retrieval system. However, even if the problems related to the varying capture conditions can be avoided, the likely mismatch between the query and collection images in terms of the viewpoint from which an object or a scene are captured still remains the main obstacle for the successful practical adoption of QBE-based image retrieval. This obstacle is particularly critical since it makes the retrieval performance inconsistent with a user's expectations. For example, a user may expect a good retrieval result given a query image of a high quality. However, a high-quality query may perform worse than a low-quality one if the object in the high-quality query is captured from a very different viewpoint from that for the collection images. This problem is illustrated in Fig. 6.1 using a set of queries which are visually similar. The frames extracted from a video clip about a landmark of the Oxford University are used to query the images in the Oxford building dataset [6]. As shown in Fig. 6.1, the retrieval performance varies greatly if different video frames are taken individually as queries, although they all show the same object and are visually similar.

While the example in Fig. 6.1 is used to illustrate the query-collection mismatch problem as the main reason for unreliable QBE-based image retrieval, this example also reveals a potential effective solution to this problem. Multiple images of the same object that are characterized by different capture conditions and viewpoints could, namely, be aggregated together in order to extract the

[2] http://www.google.com/mobile/goggles/
[3] http://www.tineye.com/
[4] http://www.bing.com/images

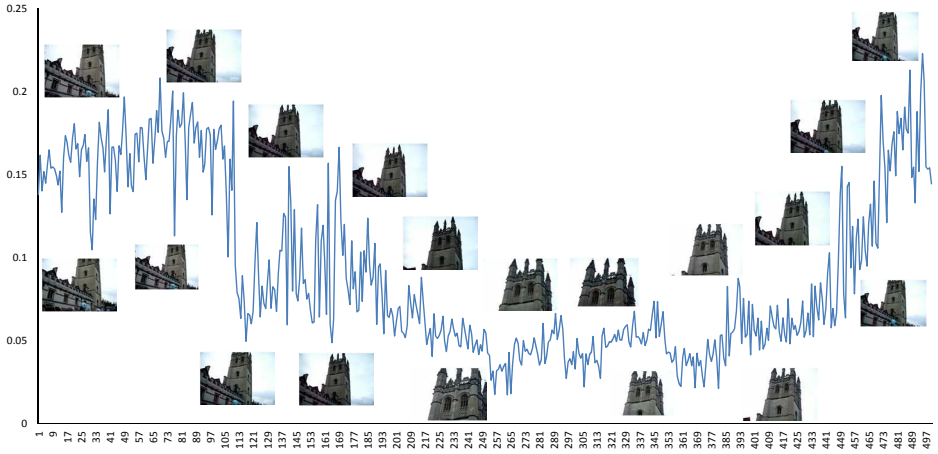

Figure 6.1: *Illustration of the variance in the retrieval performance (average precision) using different captures of the same visual object as query. The query images are extracted from the segment beginning at 7s and ending at 27s of the video found at http://www.youtube.com/watch?v=ehPaPXaxQio. The vertical axis is the average precision.*

information for creating a more robust representation of the query object, a representation that is more complete than if any of the individual images are used as query alone. Although multiple images of the same object can be collected in various ways, a video capturing the object provides the most intuitive way to generate such a complex query, as it removes the need for the user to decide about the type and number of images to take for the same object. We therefore refer to this promising solution to the query-collection mismatch problem further as *video-based image retrieval* (VBIR).

Video-based image retrieval is also regarded as a useful extension to query-by-example video retrieval [89], which uses video query to retrieve videos in the collection. Compared with QBE-based video retrieval, VBIR can provide an alternative way to satisfy users in many application scenarios. First, although a video contains more information than a single image, it may be more convenient for users to browse image search results than video search results in a handheld small screen device. Furthermore, video browsing suffers from adaptation problem in small screen devices. Second, the metadata accompanied with or the web pages containing an image are usually more descriptive and informative for users to understand the contained object than that for a video.

In this chapter, we investigate the potential of the VBIR concept for improving QBE-based image retrieval. Due to the convenience of video capture in a mobile search scenario and the high practical importance of successfully realizing QBE-based image retrieval there, we focus on this particular scenario. As a

consequence, we not only propose a method for improving the quality of retrieval results using the VBIR concept, but also a method for improving the retrieval efficiency under this retrieval concept.

The chapter is organized as follows. After reviewing the related work and positioning our contribution with respect to it in Section 6.2, we provide in Section 6.3 an overview of our proposed VBIR framework. Then, in Section 6.4, we describe the key components of the framework in more detail, which is followed in Section 6.5 by the description of the proposed algorithms for improving the retrieval efficiency. The experimental evaluation of the proposed approach is presented in Section 6.6, followed by the conclusions and prospectives for future work in Section 6.7.

6.2 Related Work

QBE-based image retrieval is one of the first retrieval paradigms introduced in the field of multimedia information retrieval, and has been extensively studied already for two decades [95][23]. Since recently, it has gained increasing attention due to a number of successful commercial applications built on this retrieval paradigm. For example, TinEye released a reverse image search engine to retrieve web pages containing the near-duplicates of the query image. Bing Image Search released a new feature called "Find more sizes", which allows users to retrieve different sizes of images that are near-duplicates of the query image. Particularly addressing the challenge of image retrieval in a mobile use scenario, Google Goggles was developed to allow search for information using an image captured by a mobile phone. The retrieval mechanisms underlying these applications are mostly based on image representation and matching using SIFT features [61] and the concept of bag-of-visual-words [93][77] derived from these features.

While the development of SIFT-based image representation solutions has been remarkable over the past several years, it is unrealistic to expect that this development could lead to a perfect image representation for any retrieval use case. Therefore, the idea behind the VBIR concept proposed in this chapter is not to work towards an improved feature-based image representation, but rather to put the currently available and imperfect features into a good use, by incorporating relevant auxiliary information. Working in this direction, Yang et al. [124] proposed to incorporate the visual context of the object captured by the query image to enrich the visual query representation and in this way improve the relevance of the retrieved images. In this chapter, we enrich the query representation by drawing benefit from the information contained in the multiple frames of the query video in order to compensate for the deficiencies of a single-image query.

The proposed VBIR approach is partially related to several recent works in the field. In [92] Sivic et al. proposed an application to retrieve the shots in a given video similar to the query shot in terms of the object of interest captured in the query shot. There, feature tracking is used to identify the object of interest

in the frames of the query shot. Then, the search is performed by using different frames in the query shot individually as query images, after which the partial results are aggregated into the final search result. Compared to [92], we also use a video clip as query, but target the general (unconstrained) image search problem. Furthermore, we explicitly address the problem of improving the query representation by searching for the most stable feature points and by constructing query expansion using synonyms. Finally, we propose a comprehensive solution to VBIR including the postprocessing of the search results using reranking and taking into account the issues related to the implementation efficiency in view of the targeted mobile image search scenario.

Query expansion using an automatically constructed synonym is a well known technique in information retrieval [67] and has been utilized in [21][106] for improving image retrieval performance. While in [21][106][66] the synonym is learned from the database, in our approach the synonym is learned from the query video, which is more effective and also more adapted to the user's current search intent. Although in both [98] and our approach video is used as the context to improve image retrieval, our proposed approach is different from [98] in that we use the video directly as query, while in [98] the video context is used to learn the parameters offline for a domain-specific image feature representation. As a result, the developed approaches are entirely different.

The research on efficient implementation of image retrieval systems has mainly focused on searching for efficient image representation features [11] and on efficient implementation of existing features [35]. Wagner et al. [111] proposed to utilize the feature tracking results to reduce unnecessary detection of feature points for an efficient implementation of image search on mobile phones. Our proposed priority queue based feature description addresses this efficiency problem in a different way, namely by optimally using the limited time budget. The visual word quantization is often realized using fast approximate nearest neighbor search [77][72]. However, the feature points are mostly quantized independently. To exploit the redundancy across the frames in a video, we propose the cache-based bi-quantization to quantize the feature points jointly to further reduce the time cost.

6.3 Video-based Image Retrieval

A system overview of our proposed VBIR framework is illustrated in Fig. 6.2. The considered use scenario is that users first capture a video clip on the object of interest using their mobile devices and then submit it to the VBIR system to retrieve images or other information relevant to the object of interest. The retrieved images can be browsed by users, or the metadata associated with these images can be presented to help users understand the observed object.

When a query video is submitted to the system it is decoded into a frame sequence and the local features, such as SIFT [61], are extracted from each frame.

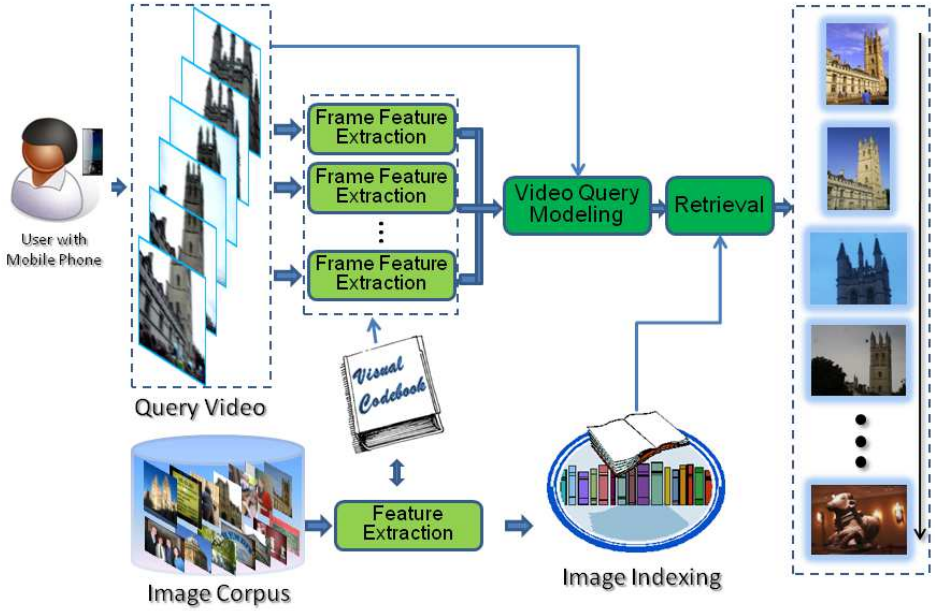

Figure 6.2: *System overview of the proposed video-based image retrieval (VBIR) framework.*

The features are then quantized into visual words based on already built visual codebook. The codebook is built in the same way as in a typical QBE-based image retrieval system, namely using Approximate Kmeans algorithm [77]. Then the SIFT features are mapped onto visual words using hierarchical Kmeans tree algorithm [72]. The generated visual words for all the images in the collection are indexed using the inverted file structure [67]. Furthermore, frame-level visual words aggregated over video frames are used to derive an improved query representation. Finally, we retrieve the images from the collection based on the improved query representation and present the results to users. In the following section, we focus on the core of our system, where the improved query representation is derived from multiple video frames and used to improve the retrieval results.

6.4 The Proposed Approach

Given the visual words extracted from all video frames and the temporal structure information in the query video, we need to appropriately process the video query and design a retrieval model in order to be able to draw maximum benefit from the rich information contained in the query video. Fig. 6.3 illustrates the flowchart of the proposed VBIR approach zooming in on the query processing and retrieval

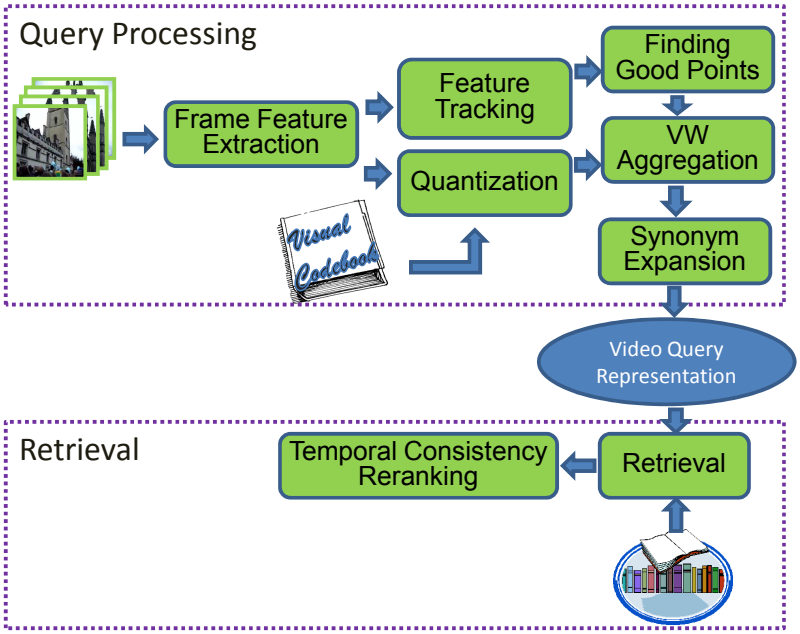

Figure 6.3: *The flowchart of the proposed VBIR approach zooming in on the query processing and retrieval steps.*

steps. In the query processing step, we first perform feature tracking among the detected SIFT feature points over adjacent video frames and then find "good" points, which are stable and therefore able to represent the query well. After that, the good points are aggregated into a histogram to obtain a first improved query representation. This representation is then further expanded based on the mined synonyms. In the retrieval step, temporal consistency reranking is introduced to further refine the search result obtained by a general image retrieval model based on the expanded query. In the following, we will describe the elements of the flowchart in Fig. 6.3 in more detail.

6.4.1 Corresponding SIFT points among frames

First, we track the SIFT points over all video frames to construct the corresponded point sequences. Here the corresponded point sequence is defined as a sequence composed of the SIFT points in adjacent frames, which can be corresponded by tracking. The construction of corresponded point sequences is performed as follows. For each pair of adjacent frames in the query video, we firstly track the SIFT points detected in the previous frame using Lucas and Kanade optical flow algorithm [62] implemented in OpenCV and modified using image pyramids [14]. Then, the tracked positions in the subsequent frame are further aligned to the

detected SIFT points. Specifically, we find all the SIFT points which are not more than one pixel far from the tracked positions as the tracked SIFT points. The process is repeated for each pair of adjacent frames in the query video to produce many corresponded point sequences, each of which comprises a sequence of tracked SIFT points across video frames.

All the corresponded point sequences obtained using the procedure described above comprise the set $\mathcal{S}$. $\mathbf{S}(p_i^j)$ denotes the point sequence comprising SIFT point p_i^j, the j^{th} SIFT point in the i^{th} frame. For the convenience of implementation, those SIFT points which cannot be corresponded are also added into the point sequence set. Consequently, each of these sequences comprises only one SIFT point.

6.4.2 Finding good points

We assume that a good point that is reliable for retrieval should have the following properties. First, it can be tracked and corresponded in multiple adjacent frames, which states that it is stable and clearly identifiable. Second, it should gravitate towards the center of the frame, which is due to our observation that people usually tend to put the object of interest in the center of the frames when capturing a video, so the central points are more likely to be related to a users' search intent.

Based on the above assumptions, we design a set of criteria to evaluate the goodness of points. For each point p_i^j, its corresponded point sequence is denoted as $\mathbf{S}(p_i^j)$. Then, the goodness of p_i^j is defined by Eqn. (6.1) as a combination of two terms, the *stableness* term and the *center-awareness* term,

$$G(p_i^j) = \alpha \times \frac{Len(\mathbf{S}(p_i^j))}{FrameCount} + (1-\alpha) \times Cent(p_i^j). \tag{6.1}$$

Here, α is a parameter to control the respective contributions from the two terms, and $FrameCount$ is the number of frames in the query video, which is used for normalization. $Len(\mathbf{S}(p_i^j))$ denotes the number of frames being tracked in the point sequence $\mathbf{S}(p_i^j)$ to represent the stableness of the point. The center-awareness term $Cent(p_i^j)$ is defined to reflect the assumption that the object near the center of the image is of more importance. Considering the occasional departures of intended objects from the central image area, we use the average distance of all the points in the tracked sequence to represent the center-awareness of each point in the sequence. The center-awareness of point p_i^j is defined as,

$$Cent(p_i^j) = -\frac{\sum_{p \in \mathbf{S}(p_i^j)} d(p,c)}{Len(\mathbf{S}(p_i^j)) \times d(0,c)}. \tag{6.2}$$

Here, d denotes the distance from point p to the frame center c, and $d(0,c)$ represents the distance from the origin of the frame to the center.

After the goodness of the points has been computed, we select those points with a goodness value larger than a threshold as good points, which will be used to construct the query model, as will be explained in the following sections.

6.4.3 Aggregating visual words

Given the good points selected in all the frames in the query video, we now aggregate them to construct an improved query model as a bag of corresponded visual words. For efficiency reasons, the temporal information of the points is not used in the query representation. However, we noticed that the temporal information may be important to further improve the retrieval result. Hence, we incorporate the temporal information into the reranking process described in Section 6.4.5 for a trade-off between the retrieval effectiveness and efficiency.

The query video is represented as a histogram, denoted as $\mathbf{q}$, where each bin q_i corresponds to a visual word w_i in the vocabulary. Then, for each visual word, we aggregate its occurrence in all frames, divided by the number of frames in the query video, as the value in the corresponding bin of the query histogram. Representing the query as an aggregated histogram is a convenient way to take into account all the appearances of the query object in different frames with variations including scales, viewpoints, and lighting. It utilizes the redundancy in the video to achieve a comprehensive representation of the object of interest captured by the query video. In addition, compared with that of fusing the retrieval results using different video frames as query, which requires multiple scan of the database [92], the aggregation of visual words into a single query representation makes the retrieval process more efficient.

Even though the aggregated visual words already contain rich information that should be sufficient to enable a more reliable retrieval compared to a single-image query case, we will show in the following that reliability could be improved even further, by mining the video for query synonyms to further expand the query representation.

6.4.4 Synonym expansion

While SIFT features are generally effective for image retrieval, different SIFT descriptors can still be extracted for the same object patch in different images, due to which similar images with large variations cannot be matched well. The visual word quantization, which is used to improve the retrieval efficiency and scalability, makes this problem even more severe since the quantization error brings additional obstacle for matching the image patches.

One of the advantages of a video compared to a single image is that it may contain a wide range of different appearances of the same object. This redundancy provides useful information for deriving the relations between the features extracted in different frames. Stavens et al. [98] used such information to learn the parameters for feature description. In this chapter, this information is utilized

to construct the synonym relations among visual words to partially address the imperfections due to visual word quantization.

For each visual word w_i, its term count in all frames of the query video is denoted as tc_i, and the number of points in a corresponded point sequence $\mathbf{S}_k$ being quantized as w_i is denoted as $tc_i(\mathbf{S}_k)$. Then we can construct an affinity matrix M with the element m_{ij} defined as follows,

$$m_{ij} = \frac{\sum_k \min(tc_i(\mathbf{S}_k), tc_j(\mathbf{S}_k))}{tc_i}, \tag{6.3}$$

with the diagonal elements set to zero.

The affinity matrix is then used to generate a contextual histogram from the aggregated query histogram so that the term counts of synonymous visual words can boost each other to alleviate the problem of quantizing similar feature descriptors into different visual words.

The contextual histogram is generated as,

$$cq = M \cdot q. \tag{6.4}$$

This histogram is then combined with the aggregated query histogram into the new query representation,

$$q_{new} = \beta q + (1 - \beta) M \cdot q. \tag{6.5}$$

Using the new query representation, we can construct the vector space model based on the standard *tf-idf* scoring function known from text information retrieval to compute the similarity between the query video and images in the collection:

$$q_v = q_{new}. * idf. \tag{6.6}$$

Here the operator $.*$ stands for element-wise vector multiplication, while idf is a vector where idf_i represents the idf (inverted document frequency) of the visual word w_i. Then the cosine similarity function can be employed to compute the similarity between the query q_v and image features.

6.4.5 Temporal consistency reranking

While many frames in the query video have been employed to achieve a robust query representation, the noisy information spread in the frames may also get aggregated to produce an amplified negative effect on the query quality. Hence, to suppress this negative effect while keeping the advantages of visual word aggregation, we propose a reranking approach to adjust the search result achieved in the above steps by taking the temporal consistency of the visual content into consideration.

The reranking approach is based on our assumption that the false matches between the query video frames and the database images should not be consistent

among adjacent video frames. In other words, since adjacent frames usually do not exhibit great changes in their appearance and all contain the object of interest, the similarity scores computed between a relevant image in the collection and adjacent frames in a video should not change greatly. However, for a mismatch, a high similarity score obtained on one video frame, e.g. due to noise in feature representation and capture conditions, will most likely be followed by a low score on the next video frame. In view of this, we choose to rerank the images in the top of the results list based on the temporal consistency information.

For each image I_i in the top of the results list, we compute the similarity scores between that image and all frames in the query video based on the vector space model with *tf-idf* weighting, denoted as $v(I_i, F_k)$, where F_k represents the k^{th} frame in the query video. Then by regarding $v(I_i, F_k)$ as a function of k, we can compute the gradient of the function as

$$g_i^k = v(I_i, F_{k+1}) - v(I_i, F_k). \tag{6.7}$$

The absolute values of the gradients are then averaged to reflect the temporal consistency of the matching scores for adjacent frames:

$$\tilde{g}_i = \frac{\sum |g_i^k|}{FrameCount}. \tag{6.8}$$

The average gradient is then combined with the similarity score computed in Section 6.4.4 to obtain a new reranking score for the top ranked results,

$$r_i = -\tilde{g}_i + \gamma \bar{r}_i, \tag{6.9}$$

where $\bar{r}$ is the initial ranking score.

We noticed that some of the query videos are highly dynamic due to camera shake. For such a query video, all the images in the database may have a high average gradient, which implicitly increases the impact of temporal consistency on reranking. Actually, for the highly dynamic videos we want to decrease the contribution of temporal consistency to reranking since even for a positive sample it cannot achieve a low average gradient in such cases. We use the mean of average gradients of the top-ranked images as the measure of the dynamics degree of the query video, which is then used to weight the average gradient term to achieve a new reranking function. In this way, the expression in Eqn. (6.9) can be modified as

$$r_i = -\frac{\tilde{g}_i}{\frac{1}{N}\sum_{i=1}^{N} \tilde{g}_i} + \gamma \bar{r}_i, \tag{6.10}$$

where N is the number of top-ranked images to be considered in reranking.

6.5 Efficient Implementation

A naïve implementation of the proposed VBIR approach may be inefficient, leading to a slow query response. Hence, to make the proposed approach applicable

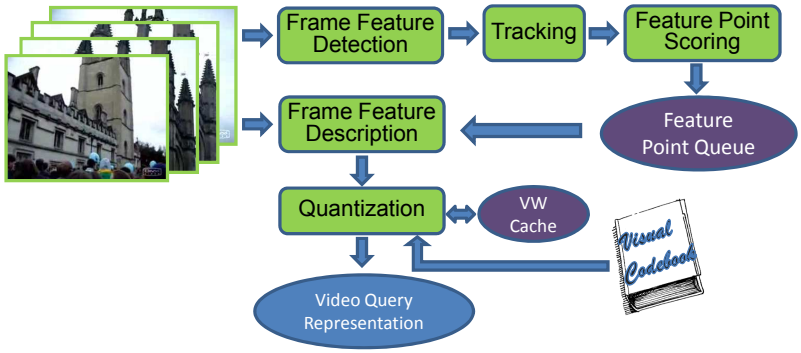

Figure 6.4: *The flowchart of the proposed efficient VBIR implementation.*

in a real-life mobile use scenario, where the realtime query response is required, we propose a pipeline described in this section to further improve the efficiency of the VBIR framework introduced before.

The proposed implementation is based on the client-server architecture. Users capture a video on the client computer or a mobile device and then upload it to the server, which is responsible to process the video and retrieve relevant images. The first issue to be considered is the network transferring. Based on our experiments, transferring a 6s 10fps 320×240 video clip over a 3G network will cost about 1.1s. In other words, it costs on average 18ms to transfer one video frame. Hence, by adopting the progressive uploading or streaming, the video uploading may become realtime, which means that the whole query video can be transferred to the server in a short time after the video capturing is completed.

To identify the computational bottleneck, we analyze the computational cost of the components of the proposed VBIR framework[5]. The entire query processing part costs about 650.82 milliseconds for processing a video frame. There, the most time-consuming component is the SIFT feature extraction and quantization, which costs about 345.23ms and 255.43ms, respectively. The SIFT feature extraction contains two separate processes, interest point detection and description, and they cost about 99.11ms and 243.01ms, respectively. From these results, it can be observed that the feature description and quantization cost in total about 76.59% of the computation of query processing and therefore jointly form the bottleneck of query processing in our VBIR framework, The retrieval step, on the other hand, is efficient. It only takes 1.28s to handle one query.

Based on the above analysis, we propose an efficient VBIR implementation, as illustrated in Fig. 6.4. Since the processing of different frames is independent of each other, the video query processing can easily be parallelized. We maintain a thread pool comprising three threads, and for each input video frame, a feature

[5]The experiments about the computational cost in this chapter are performed on a workstation with two dual-core Intel Xeon 2.67GHz CPUs and 12GB memory.

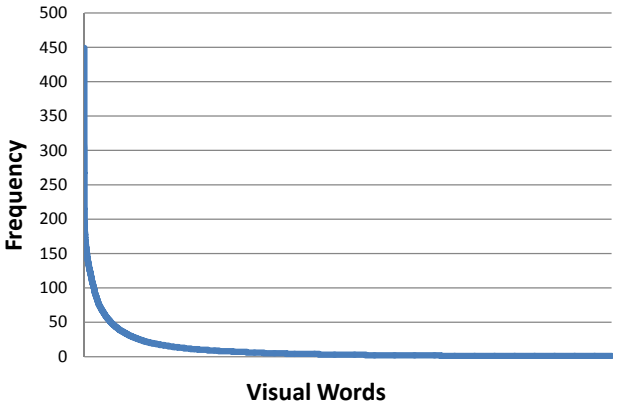

Figure 6.5: *The frequency of visual words in the query video All_Souls_v1.*

point detection thread is created and added into the thread pool. Since the feature description and quantization are time-consuming, we develop a priority queue based mechanism, so that the most important feature points can be processed in a limited time budget, as will be introduced in Section 6.5.1. To further reduce the quantization time, we rely on the fact that the feature points in adjacent frames are often similar and correspond to the same visual words. Hence, we develop a cache-based bi-quantization algorithm for speed-up, as presented in Section 6.5.2.

6.5.1 Priority queue based feature description and quantization

In a realtime image search system, the search results should be returned quickly after the user has finished uploading the query video. Hence, there will be only a limited time budget available for the query video processing. In such a limited time, it may be difficult to process all detected feature points in the video. Hence, we will maintain a priority queue to keep all the detected feature points for which the description has not been extracted. For each frame, after tracking, the newly detected feature points will be enqueued and the priority of points in the former frames will be updated based on Eqn. (6.1). The feature description thread will continuously fetch the feature points from the queue for processing, until the queue is empty or the time budget has been used up.

6.5.2 Cache-based bi-quantization

The proposed cache-based bi-quantization algorithm is motivated by the fact that there is a *local consistency* in the visual word quantization for adjacent frames in a query video. Since the adjacent frames are normally very similar to each other, the descriptions of the feature points in adjacent frames also tend to be similar to each other and the quantized words would be identical. Fig. 6.5 shows

Algorithm 1 Cache-based bi-quantization

Require: a high-precision quantizer Q_h, a low-precision quantizer Q_l, and N_l which is the frame interval for cache refreshing.

Initialization Set cache $\mathcal{C} = \varnothing$.
for Frame $F_i = F_1$ to F_N **do**
  for all Feature point P_j in F_i **do**
    Quantize P_j into visual word W_j using Q_l: $W_j = Q_l(P_j)$;
    if $W_j \in \mathcal{C}$ **then**
      continue;
    else
      $W_j = Q_h(P_j)$;
      $\mathcal{C} = \mathcal{C} \cup \{W_j\}$;
    end if
  end for
  if i%N_l==0 **then**
    $\mathcal{C} = \varnothing$;
  end if
end for

the occurrence times of each visual word into which the feature points in a query video are quantized. We can see that 6.60% visual words occur more than 40 times in the query video All_Souls_v1, which corresponds to 50% feature points.

To utilize the local consistency, we propose a cache-based bi-quantization algorithm, as shown in Algorithm 1. Since the visual word quantization is normally performed by using approximate nearest neighbor search, such as k-d trees [72], we can adjust the search parameters to trade-off the precision and the time cost. In our approach, we built two quantization methods. One is Q_h, slow but with high precision, and the other one is Q_l, which is fast but with a low precision. For each feature point, we firstly use Q_l to get a rough quantization with a small time cost. To verify the reliability we check whether the quantized word has appeared in the cache. If so, it should be a reliable quantization. If not, we further achieve a reliable quantization using the high-precision quantizer Q_h. For every N_l frames, the cache will be cleared and refreshed to maintain the locality. In this chapter, we simply set $N_l = 20$.

6.6 Experiments

6.6.1 Experimental setup

To set a benchmark for video-based image retrieval and allow for comparison of other methods with the approach proposed in this chapter, we first chose the publicly available Oxford building dataset [6] as the image collection. Then, as explained in Section 6.6.5, we also expanded our investigation to a larger, web-scale image collection for the purpose of a more comprehensive evaluation of

the algorithm performance. The Oxford building dataset comprises 5062 images crawled from Flickr[6] using 11 landmarks of Oxford University as queries. To collect the query videos for our experiments, we used the 11 landmark names as query to search for suitable videos in YouTube. Finally we obtained 15 videos for 8 landmarks, while for the other 3 landmarks we were not able to find any relevant videos. Since not all the parts in these videos are about the corresponding landmarks, we selected those segments exactly describing the landmarks as query video clips in our experiments. Consequently, 25 video clips were selected as queries, which are summarized in Table 6.1. The ground-truth corresponding to each landmark in the Oxford building dataset is still used as the ground-truth for the video-based image retrieval experiments. The key frames in the query videos and the images in the database were down-sampled to 300*300 by preserving the aspect ratio and then SIFT features were extracted. A 100K visual vocabulary was constructed using Approximate Kmeans [56].

The Average Precision (AP) is used to evaluate the retrieval performance. AP is defined as the average of the precisions computed at all recall levels. The Mean Average Precision (MAP) is the average of the APs across all queries.

6.6.2 Performance comparison

We implemented QBE-based image retrieval (QBEIR) as a baseline to be compared with the proposed VBIR concept. Specifically, we used each frame in the query videos individually to query the image database to simulate QBE-based image retrieval. The average performance (QBEIR_E) with standard deviation and the best possible performance (QBEIR_B) for each query video using different frames as query are illustrated in Fig. 6.6. The methods in [92], which fuse the retrieval scores by using each frame individually as query are also implemented and used for comparisons, including fusion by summing all scores (OLG_S) and fusion by taking the maximum (OLG_B).

We can see from Fig. 6.6 that QBE-based image retrieval suffers from a dramatic performance variation, which demonstrates its insufficient reliability. Intensive camera motion, e.g., zoom in/out in HertFord_v1, and large changes of light conditions caused by different shooting angles in Christ_Church_v1 and Radcliffe_Camera_v1 cause that the object of interest is described at a broad range of capture conditions, which can only in part match the conditions at which collection images have been captured. This causes large variation in the retrieval performance if video frames are used individually as query. However, such information can be put into a good use to improve the retrieval performance by using the whole video clip as query, as proposed in this chapter.

The performances of VBIR and QBE-based image retrieval is compared in Fig. 6.6. We can see that the performance of VBIR is significantly better than the expected performance of QBE-based image retrieval. The improvement was

[6]http://www.flickr.com

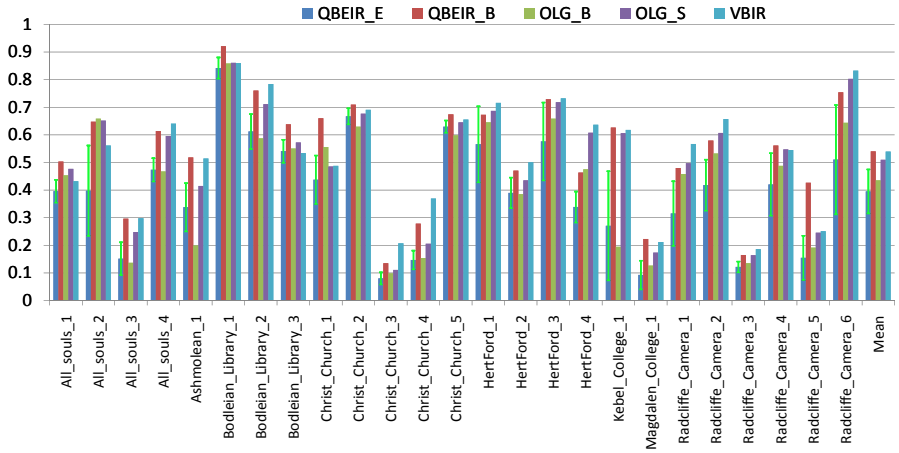

Figure 6.6: *The performance comparison of QBE based image retrieval and VBIR on the Oxford building dataset. QBEIR_E shows the mean and the standard deviation of the MAP of image retrieval using single frames in the query video, QBEIR_B is the best possible performance of retrieval using a single frame, and VBIR shows the MAP of video-based image retrieval.*

computed as 36.37% in terms of MAP. Furthermore, for 24 among 25 query videos VBIR can achieve a performance boost compared to the average performance of QBE-based image retrieval. In particular, the MAP of VBIR is even larger than the expected performance of QBEIR by a margin of the standard deviation for 19 queries. This demonstrates that the incorporation of the information contained in all video frames has a clear potential for improving the reliability of the retrieval performance. Finally, we can see that VBIR achieves a comparable result with the best possible performance of QBEIR. This means that, by using a video as query, we can achieve a result similar to the best one achievable by using an arbitrary single image as query.

The proposed VBIR approach further outperforms OLG_B and OLG_S by 23.95% and 5.81% and in 84% and 80% queries, respectively. While VBIR achieves a better performance compared to its competitors, it also exhibits a better efficiency in terms of the retrieval part. While OLG_B and OLG_S cost 5.2s to complete the retrieval part for one query, VBIR only needs 1.2s for the same task.

For those videos that exhibit significant camera motion while introducing new information such as the object at multiple scales or viewpoints, e.g., Magdalen_College_1 and All_Souls_2, incorporating multiple frames into the query representation significantly improves the retrieval performance. However, for those videos in which all frames have the same scale and viewpoint, e.g. Christ_Church_2, VBIR cannot provide a large benefit since a video in that case hardly contains more useful information than a single image. We believe, however, that when

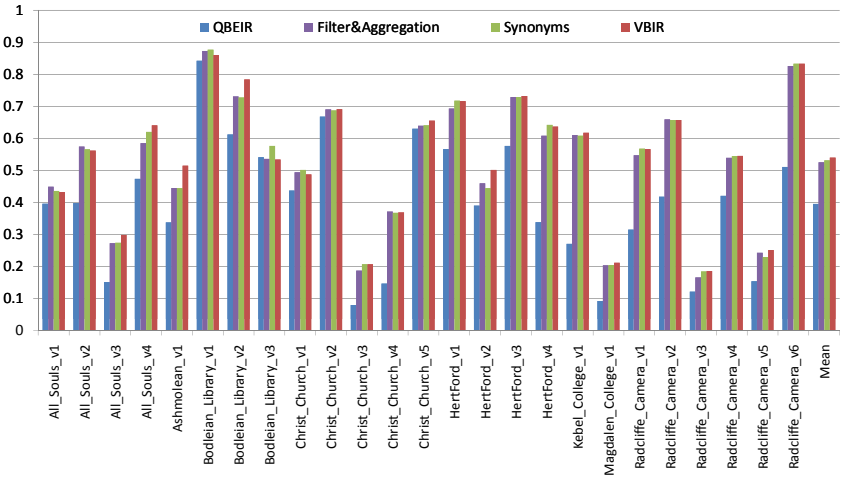

Figure 6.7: *The performance of each step in our proposed approach.*

users search for something using videos, it is realistic to expect that camera will move and zoom in/out will be deployed in order to capture more information. In this sense, VBIR is expected to boost the retrieval performance in most cases, compared to QBE-based image search.

6.6.3 Analysis of the proposed approach

To analyze the effects of each step of our proposed approach, we compare the intermediate results of the three steps, including filter&aggregation, synonyms mining, and temporal consistency reranking in Fig. 6.7. Since the result after temporal consistency reranking is just the result of the complete approach, it is denoted as VBIR in the figure.

It can be observed that by introducing the Filter&Aggregation step the performance is improved by 0.130 over QBE-based image retrieval and that the performance is boosted for 24 queries. We argue that the reasons for this effect are two-fold. First, in the "finding good points" step, the noisy SIFT points are filtered out so that they do not have a negative effect on the retrieval. For example, as shown in Fig. 6.8, the trees in the background are filtered out by our approach. Second, aggregating visual words over all frames collects the appearances of the object taken under different conditions, so that a more comprehensive representation of the query object is constructed to improve the retrieval performance. This is especially useful when a single image can only capture a partial view of the object of interest, which is likely to happen if the user stands near the object with a common camera without ultra-wide-angle lens. For instance, each frame in query Radcliffe_Camera_v2 only contains a part of the building while aggregation

Figure 6.8: *The examples of "good" feature points. The green points are good points and the red ones are those being filtered out.*

brings us a full view of the object and therefore a better retrieval performance.

By incorporating the synonyms mining, the performance is further improved by 0.006, as shown in Fig. 6.7. Moreover, we can see that for a large majority of queries, introducing the synonyms boosts the retrieval performance. Among them, we observe that a video with drastic camera motion, e.g., zoom in/out, such as HertFord_v1, and light condition changes like Radcliffe_Camera_v1 and Christ_Church_v1, tends to achieve a larger improvement. Since such videos contain different views of the object of interest due to the camera motion and light condition changes, the synonyms mining can discover the correlation between the visual words under different views or scales. By including the visual word correlation into the retrieval process, the system reliability is further improved.

The temporal consistency reranking step further contributes 0.008 to the overall performance. As illustrated in Fig. 6.9, the temporal consistency assumption is verified to discriminate the positive from negative images. In such query videos, the incorporation of temporal consistency reranking improves the performance significantly. However, we also notice that for some videos the temporal consistency reranking even degrades the performance. For example, on the query HertFord_v1, the performance is degraded by 0.001 after reranking. By observing the video clip, we found that this query video contains a significant shot (dissolve) change, which breaks the temporal consistency assumption. However, we note that in a real-life VBIR system, a user-captured short video clip is not expected to contain shot changes.

6.6.4 Efficiency

While the above experiments demonstrate that the proposed VBIR approach is effective, it will cost 650.82ms to process one frame in a query video and therefore

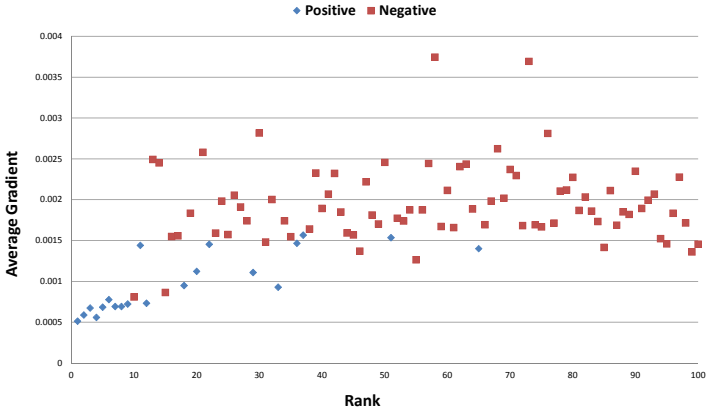

Figure 6.9: *The distribution of the average gradients for top 100 images of the query All_Souls_v2.*

it is difficult to apply in a search system where realtime query response is required. In this subsection, we will show that after adopting the proposed efficient implementation, the time can be reduced to less than 300ms, which makes the system able to process a 10fps query video in real time with three threads on a normal computer.

Figure 6.10 shows a comparison between the proposed cache-based bi-quantization and the baseline quantization approach that quantizes each feature points independently. It can be observed that under the same time budget, the cache-based bi-quantization approach can achieve a better MAP than the baseline. In other words, to achieve the same effectiveness, the cache-based bi-quantization can perform faster. Further, the MAP of the cache-based bi-quantization with Q_h being a 0.9 precision quantizer and Q_l a 0.5 precision quantizer is 0.5404 and better than that of the baseline quantizer with precision 0.9. However, the quantization time is reduced by 32.80%, from 0.4ms to 0.27ms for one feature point. Hence, in our experiment, we used the cache-based bi-quantization with 0.9 precision Q_h and 0.5 precision Q_l.

From Fig. 6.11 we can see that the cache-based bi-quantization without cache refresh achieves the highest speed but the lowest MAP, which validates the locality of the visual words consistency and demonstrates the necessity for cache refresh. Based on the result, we can set the refresh interval to a moderate size, e.g., 20 frames, to achieve a trade-off between the effectiveness and efficiency.

To study the relationship between the MAP and the time cost of the priority queue based feature description and quantization, we illustrate in Fig. 6.12 the respective MAP of filtering out different percentages of feature points to reduce the time cost. We can see that by filtering a small amount of points (less than 30%) the performance even improves over that using all feature points. This

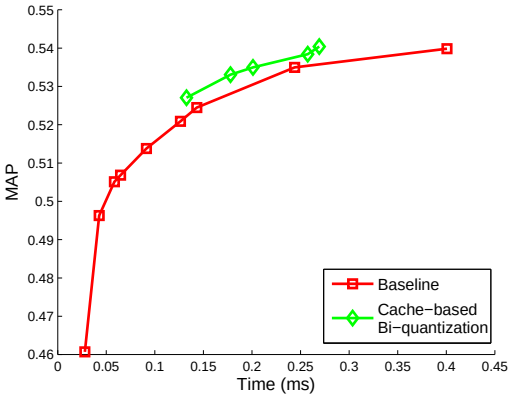

Figure 6.10: *The performance comparison between the cache-based bi-quantization and the baseline quantization method.*

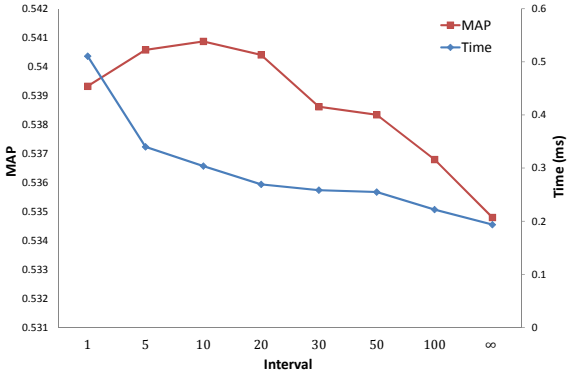

Figure 6.11: *MAP and time for cache-based bi-quantization with different intervals for refreshing the cache.* ∞ *means no cache refresh.*

demonstrates the effectiveness of the proposed feature filtering step described in Section 6.4.2. By filtering 70% feature points, we can still achieve 0.4841 MAP, which improves 22.3% over QBEIR. But the total time for query processing is reduced to 283.12ms, which can be completed in realtime using three threads.

6.6.5 Experiment on a large-scale dataset

To further demonstrate the effectiveness of the proposed VBIR approach, we performed another experiment on a large-scale dataset. The videos of Oxford university buildings we crawled from YouTube were still employed as queries, but the database was composed of not only the images in Oxford building dataset,

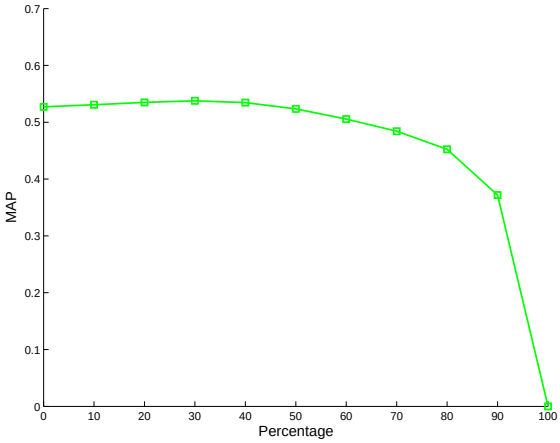

Figure 6.12: *MAP for different percentages of visual words to be filtered out.*

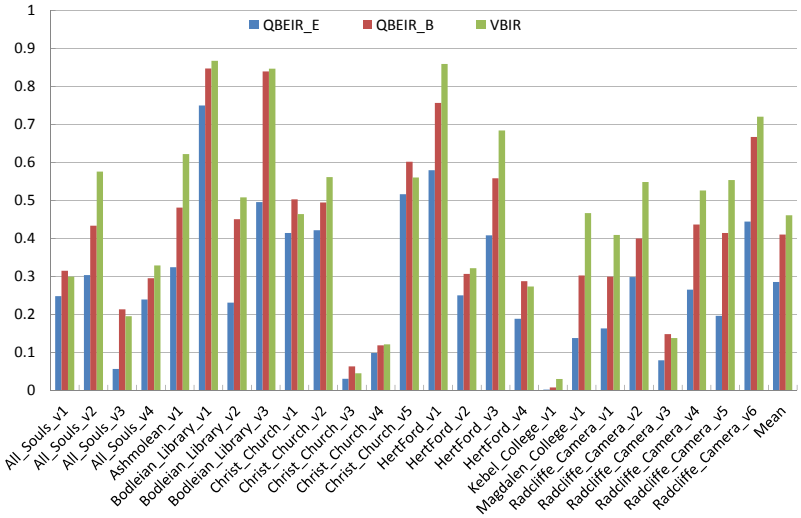

Figure 6.13: *The performance of VBIR on a large-scale dataset comprising 1M images.*

but also one million images collected from Flickr. Finally the database comprises totally 1,004,834 images. The performance of our proposed VBIR approach, and the expected and the best performance of QBE are shown in Fig. 6.13.

By comparing Fig. 6.13 and Fig. 6.6 we can see that when the database scales up, the retrieval performance of QBE drops significantly. Specifically, the MAP of QBEIR_E decreases from 0.396 to 0.285 when the scale of the database increases from 5K to 1M. This shows that the QBE-based approach is less robust

and less scalable than VBIR, which still achieves 0.461 MAP for 1M database. Moreover, we found that on the large-scale dataset, VBIR even outperforms the best possible performance of QBE (QBEIR_B), which further demonstrates the effectiveness and reliability of VBIR.

6.7 Conclusion and Future Work

We proposed in this chapter a new image search framework that we refer to as the *video-based image retrieval* (VBIR) framework. VBIR makes it possible to search for images and related information using a short video clip taken on the object of interest as query. The approach underlying the proposed framework includes mining of the useful information from all frames of the query video and using this information to refine the query representation and in this way improve the retrieval performance. The experimental results show that video-based image retrieval significantly improves the retrieval reliability over that of using a single image as query.

Since video-based image retrieval as a paradigm is particularly of importance for the mobile use scenario, where visual queries are captured using a mobile device, we also addressed the efficiency of VBIR framework implementation to make it deployable in a practical (mobile) use case.

We envision four main directions for future work building on the insights presented in this chapter. First, we will focus on utility aspects of the VBIR concept and further investigate the expected properties of the acquired query videos and their relation to a users search intent based on the typical user behavior when capturing videos for VBIR. Domain-related insights collected here will help us to further improve the framework from the design and implementation perspective. Second, we will expand the VBIR concept to investigate other possibilities for drawing benefit from the rich information contained in a video to improve the effectiveness and efficiency of query representation. One possibility is to generate a 3D object model from the query video clip and then use that to retrieve images in the database. The other is to discover useful information from the aspects typical for the video nature of the query, like motion patterns, to efficiently process the video query and prepare it for search. Third, the efficiency of the VBIR implementation could be further improved by relying on more efficient features, e.g. SURF [11]. To identify the fourth direction for future research, we note again that the goal of this chapter was to investigate the potential of VBIR to improve the image search performance relative to the conventional search using a single image as query. However, in order to also achieve significant improvement in the search performance in the absolute sense, a broader investigation is required involving the criteria related to the quality of the video query, and in particular the cases where the query does not optimally capture the object of interest, e.g. due to occlusion or insufficient focus. Construction of the representative sets of video queries for this purpose and identifying the possibilities to effectively cope with

Table 6.1: *Summary of query video clips used in the experiments.*

Query id	Video url	Time (s)
All_Souls_v1	http://www.youtube.com/watch?v=C1hwL-QHiec	0 - 6
All_Souls_v2	http://www.youtube.com/watch?v=V-sn0vkVYXo	77 - 86
All_Souls_v3	http://www.youtube.com/watch?v=V-sn0vkVYXo	152 - 159
All_Souls_v4	http://www.youtube.com/watch?v=V-sn0vkVYXo	260 - 280
Ashmolean_v1	http://www.youtube.com/watch?v=2g8G2XDJZZ4	6 - 11
Bodleian_Library_v1	http://www.youtube.com/watch?v=oGkHvCa1hRQ	6 - 13
Bodleian_Library_v2	http://www.youtube.com/watch?v=Mxjue1nf6oE	3 - 8
Bodleian_Library_v3	http://www.youtube.com/watch?v=XxNhfgL0nUk	28 - 33
Christ_Church_v1	http://www.youtube.com/watch?v=L3mvKQorVRY	16 - 18
Christ_Church_v2	http://www.youtube.com/watch?v=CCOMJ3boZIY	18 - 21
Christ_Church_v3	http://www.youtube.com/watch?v=o4ywV2cQ0Q4	7 - 10
Christ_Church_v4	http://www.youtube.com/watch?v=o4ywV2cQ0Q4	15 - 18
Christ_Church_v5	http://www.youtube.com/watch?v=o4ywV2cQ0Q4	195 - 199
HertFord_v1	http://www.youtube.com/watch?v=jtgRA9Abxs4	9 - 17
HertFord_v2	http://www.youtube.com/watch?v=OwxYkWwsgLE	143 - 144
HertFord_v3	http://www.youtube.com/watch?v=OwxYkWwsgLE	158 - 160
HertFord_v4	http://www.youtube.com/watch?v=OwxYkWwsgLE	168 -171
Kebel_College_v1	http://www.youtube.com/watch?v=KpuC-yj_uc0	11 - 14
Magdalen_College_v1	http://www.youtube.com/watch?v=ehPaPXaxQio	7 - 27
Radcliffe_Camera_v1	http://www.youtube.com/watch?v=C1hwL-QHiec	52 - 60
Radcliffe_Camera_v2	http://www.youtube.com/watch?v=jtgRA9Abxs4	64 - 69
Radcliffe_Camera_v3	http://www.youtube.com/watch?v=qhAVFlSwQ3c	16 - 18
Radcliffe_Camera_v4	http://www.youtube.com/watch?v=Pf6JHXhUgtg	52 - 59
Radcliffe_Camera_v5	http://www.youtube.com/watch?v=Pf6JHXhUgtg	67 - 90
Radcliffe_Camera_v6	http://www.youtube.com/watch?v=OwxYkWwsgLE	210 - 216

sub-optimal video queries is therefore an important future step in bringing VBIR to the following development stage.

Chapter 7

A Unified Context Model for Semantic Image Retrieval

[1]

As indicated in the previous two chapters, content-based web image retrieval based on the query-by-example (QBE) principle remains a challenging problem due to the semantic gap as well as the gap between a users intent and the representativeness of a typical image query. In this chapter, we take a further step towards solving this problem by integrating rich query-related contextual information into an advanced query model. We consider both the local and global context of the query image. The local context can be inferred from the web pages and the click-through log associated with the query image, while the global context is derived from the entire corpus comprising all web images and the associated web pages. To effectively incorporate the local query context we propose a language modeling approach to deal with the combined structured query representation from the contextual and visual information. The global query context is integrated by a multi-modal relevance model to "reconstruct" the query from the document models indexed in the corpus. In this way, the global query context is employed to address the noise or missing information in the query and its local context, so that a comprehensive and robust query model can be obtained.

[1] This chapter was published as: Linjun Yang, Bo Geng, Alan Hanjalic, Xian-Sheng Hua, "A Unified Context Model for Web Image Retrieval." ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 8, No. 3, 2012 [125].

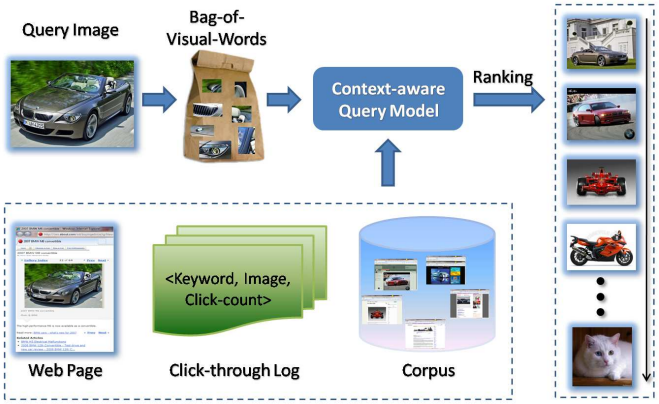

Figure 7.1: *Overview of web image retrieval using local and global query context.*

7.1 Introduction

While content-based image retrieval (CBIR) based on the query-by-example (QBE) paradigm has been extensively studied already for decades [95][23], it has recently regained increasing attention from both the industrial and research communities. A couple of interesting applications based on scalable QBE-based image retrieval have even been deployed on the web and/or on mobile clients. For example, the TinEye image search engine allows users to discover where the query image comes from, in which context it is used, whether it was modified or whether higher-resolution versions of the query image are available. Furthermore, users can use Google Goggles or SnapTell to search via their mobile phones for information related to objects and scene they see by just taking one picture and sending it to the corresponding service. Other representative applications built on the QBE-based image retrieval paradigm include automatic image annotation [116] and image advertising [68], where the objective is, respectively, to find similar images for annotation propagation [116] or to retrieve an advertisement related to the user-clicked image [68].

One of the main reasons for the revived interest in the QBE-based image retrieval lies in the advent of the bag-of-visual-words concept, which provides a powerful visual representation for effectively and efficiently computing the similarity between two images [93][77]. However, similar image search, which targets the retrieval of not only duplicates and near-duplicates but also semantically related images is still a challenging problem due to the so-called *intention gap* [139][136] and *semantic gap* [23][95][96]. Numerous approaches have been proposed to address these difficulties. Probably the most well-known of these approaches is the *relevance feedback* [142][85], which tries to bridge the two gaps through collecting additional information from the users in an iterative relevance specification procedure. It has been shown, however, that users are usually reluctant to provide

feedback to the system in real-life web search applications [67].

An alternative solution to the problem described above could be to rely on the analysis of the contextual information related to the query image. An image never appears in isolation. In the image capturing stage, the metadata such as the exposure time, the ISO, the time when it is captured, and the GPS coordinates are associated with the image to indicate the context in which it is captured. When an image is shared on the web, it is usually accompanied by some textual information (e.g. captions, comments, tags, surrounding text) in web pages associated with the image to explain the meaning of the image or its role in a broader (say, story and communication) context. These web pages can be found during the crawling stage and recorded for the later use. As these contexts are directly associated with the query image, we refer to them jointly as the *local query context*. Clearly, the information derived from the analysis of the local context of an image can be used to enrich the image representation beyond its visual (pixel-level) content.

In addition to the above, another important source for deriving the local context of a query image are the click-through logs from search engines, and in particular from the keyword-based image search engines, such as Google, Bing and Yahoo! image search. From the click-through logs, the relationship between the text keywords and images can be mined, and the corresponding keywords can be assigned to images as implicit annotations. The mined keywords for an image can be applied to enrich the existing (explicit) textual representation of the image so that a more comprehensive and robust query representation can be formed.

The local query context may, however, be insufficiently informative or noisy in some cases, like for instance, when the content of an image and the text in the corresponding web page do not (completely) match each other. Relying on the local context in such cases may even degrade the image retrieval performance. Fortunately, this problem can be alleviated by simultaneously taking the *global context* of the query image into consideration. The global query context is "hidden" in the entire data corpus consisting of all images on the web and the associated web pages. As such, the global query context can be regarded as a knowledge-base, from which the text and the visual content can be mutually interpreted and the co-occurrence patterns of textual and visual words can be mined. In this way, for example, the missing or noisy local context information could be "reconstructed" or "filtered" by learning the textual representation for the query image from the corpus.

While the potential sources for deriving contextual information for the query image are numerous, effectively incorporating the local and global query context into the web image retrieval process is a challenging problem. In this chapter, we propose an integrated context-aware image retrieval model to address this problem. The model is derived using a language modeling approach, which builds on a unified feature space integrating the local query context, or more specifically the textual representation of the query image, and the visual representation of the query in the form of visual words. After the click-through information is processed to associate the keywords to each of the clicked images, the keyword-

image relations are added as a new field into the local query context. The global query context is further incorporated through the proposed multi-modal relevance model. We namely estimate the query model not only based on the occurrence of words in the query's visual or textual representation, but also based on the visual-textual relationship implicit in all the images and web pages indexed in the corpus. Through the multi-modal relevance model the query is actually "reconstructed" from the document models indexed in the corpus so that the problem caused by the unavailable or noisy local query context can be addressed and a robust and comprehensive query model can be obtained. An overview of our approach is illustrated in Fig. 7.1.

In Section 7.2 we first position our approach with respect to the related previous work and identify the contribution of this chapter. Then, in Section 7.3 we elaborate on the two categories of query contexts defined above. Following the description of the context-aware image retrieval model in Section 7.4, we define in Section 7.5 an advanced query model incorporating the local and global query context. The experimental results of the proposed web image retrieval approach are reported in Section 7.6, which is followed by concluding remarks in Section 7.7.

7.2 Related work

Context aware information retrieval, which takes the context of the query specification and document generation explicitly into consideration to infer a better search intention model, has been recognized as a long-term challenge in the information retrieval (IR) community [12]. Various context categories, such as the user profile, the users' everyday activity, the text surrounding the query keywords in the web page, and the click-through log have been investigated in the recent literature. Surveys on this subject can be found in [55][26].

Also in the multimedia community, context-aware multimedia retrieval has received increasing attention in recent years [63]. Sinha and Jain [91] proposed to utilize the optical context of image capturing to help learn the semantic concepts found in images. Cao et al. [16] proposed to use the time and GPS information to improve the semantic concept annotation, while Yang et al. [133] mined the contextual cues including tags and GPS to improve the keyword-based image search. In [124] the authors proposed to utilize the visual context to help improve the reliability of object retrieval.

Exploiting context in multimedia retrieval often boils down to finding effective mechanisms for fusing typically multi-modal contextual information with the content information from media items to help index and retrieve these items. We can distinguish among four general approaches to multi-modal data fusion: linear combination, latent space based, graph based, and model based approaches.

Linear combination [90][42][117][19][131][69] linearly combines the relevance scores from different modalities. However, as shown by Robertson et al. [81] the

score combination typically suffers from peculiarities that can negatively influence the retrieval performance.

The latent space based approach converts the features of different modalities into a shared latent space to unify these different modalities [141][140]. However, such methods may require a data-intensive offline training stage to learn the feature mapping from the modalities onto a unified latent space, which requires a large amount of labeled training data. Graph based approach [115] firstly builds the relations between different modalities, like for instance relations between images and text using the web page structure. Then the relations are utilized to iteratively update the similarity graphs computed from different modalities. The difficulty of creating similarity graphs for billions of images on the web makes this approach insufficiently scalable.

Model based fusion methods include relevance model based methods and reranking methods. In [53] the authors employ a relevance-based language model for cross-media retrieval using the multi-media representation of documents in the corpus. The reranking methods [121][102][126] first retrieve images based on the modality same as that of the user-submitted query, and then use the representation of documents on the other modality in the corpus to refine the initial search result.

Compared to the information fusion methods discussed above, we propose in this chapter a model-based fusion method that is more effective in combining both the local and global context for web image retrieval. Our proposed method deploys a multi-modal relevance-based language model [54] and combines the local context comprising the respective fields in the associated web pages and the associated keywords from the click-through log and the global context comprising the other images in the database. Specifically, two aspects of the proposed method make it more advanced than the related existing fusion approaches. First, we take into consideration the complex structure of the associated web pages by means of a structured retrieval model using an effective model combination strategy. Second, we integrate the learning of the mapping among modalities and the use of the mapping into a unified model, which does not need offline learning and can easily scale up. The other major contribution of this chapter lies in the conclusions we have drawn from our study of the usefulness of various context categories in web image retrieval, which could be beneficial for the future research and the deployment of context in real web image search systems.

7.3 Query context in web image retrieval

While in a QBE-based retrieval task the query input, i.e. the example image, is the most important observable entity to reflect the information need of the user, user's general preferences, the search environment and the use scenario can be said to determine the *use context*, in which the information need of the user should be satisfied. In addition to the use context, also the *query context* can be

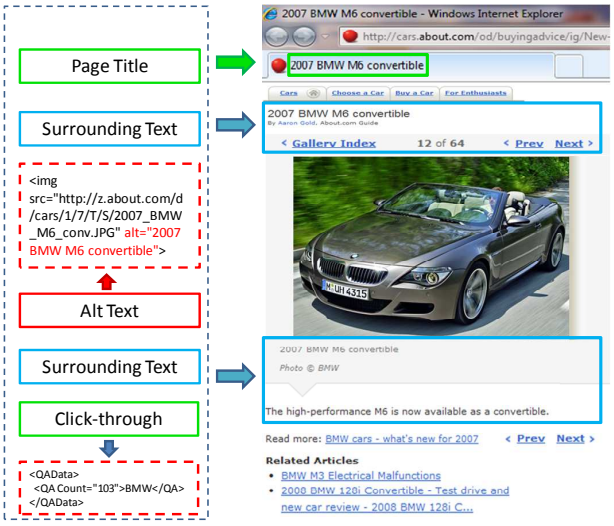

Figure 7.2: *Illustration of the proposed structure of the local query context for web image retrieval. The web page used in this figure is from http://cars.about.com/od/buyingadvice/ig/New-and-redesigned-2007-cars/2007-BMW-M6-convertible.htm.*

defined, which can help disambiguate the query and uncover the true search intent represented by the query image. Incorporating the query context into the web image retrieval process is essentially the process of expanding the actual query by linking it with different types of contextual information. Depending on the origin of this information, and in the specific case of web image retrieval, we can speak about two general categories of the query context. While the *local query context* encompasses the information that is tightly related to a particular QBE search session, the *global query context* is independent of the particular session and provides the knowledge that can be considered useful for all search sessions.

Although we focus in this chapter on the query context only, the proposed retrieval approach is general enough to incorporate other context categories. In the following sections, we elaborate on the sources of the local and global query context that we exploit in this chapter.

7.3.1 Local query context

We consider the textual information in the web page associated with the query image as the local query context. Since different text fields can be extracted from a web page, and because it can realistically be assumed that not all fields are of equal importance for the web image retrieval task, the fields will be treated separately in our approach. The fields of the local context are illustrated in Fig.

7.2 and can be summarized as follows:

- **Page title $\mathbf{q}^{pt}$**. Page title is an important field for the page author to state what the main content of the web page is about.

- **Alt text $\mathbf{q}^{a}$**. Alt (alternative) text is shown when an image cannot be displayed to a user. As such, it can be seen as a textual counterpart to the visual content of the image.

- **Surrounding text $\mathbf{q}^{s}$**. Surrounding text consists of the text paragraphs around an image in a web page. The surrounding text is in many cases semantically related to the image content and can be used to interpret the image. However, since the surrounding text can also contain information that is unrelated to the image, this field as a contextual information source can be more noisy than other fields and may therefore mislead the retrieval process.

- **Click-through log**. Click-through log is the log registering which image is clicked by users for which query. In this chapter, we specifically utilize the click-through log from a keyword-based image search engine and ignore the user identity. The click-through log can be represented as a set of tuples comprising the query, the image, and the clicked count. Based on the Query Association (QA) techniques [120][86][87], we can transform the click-through data into the set of <image, set of clicked queries> pairs by aggregating the queries for which the image is clicked for, which can be regarded as an implicitly inferred local query context and, as such, added as a new field into the textual representation of the image.

7.3.2 Global query context

The global query context can be regarded as a knowledge base from which we can mine a meaningful interpretation of the query. In this chapter, we will focus on the most straightforward, but at the same time also the most complex source of this type of contextual information, namely the collection corpus itself. The corpus $\mathcal{C}$ comprises the images on the web and their corresponding textual representations and can be used as a reference with respect to which the query image can be interpreted and the local context can be verified. In particular, the information derived from the corpus can help improve the query model estimation in cases in which the local context is unavailable or noisy, or if the user-specified query image is not representative enough to express the search intent.

Based on the above, we model the local query context as consisting of four fields, i.e., *Page title*, *Alt text*, *Surrounding text*, and *Query Association*, while the global context is represented by the *corpus*.

7.4 Context-aware image retrieval model

In this section, we first introduce the retrieval model that serves as the basis of our context-aware image retrieval approach and that is inspired by the language modeling theory. Then, we show how the introduced retrieval model can be made context-aware by incorporating the local and global query context defined in the previous section.

7.4.1 Kullback-Leibler divergence retrieval model

A retrieval approach inspired by language modeling has been widely used in information retrieval due to its sound theoretical foundation and excellent empirical performance [137]. Moreover, such an approach is flexible enough since it represents both the query and the document as language models and computes the relevance score based on the distance between the two models. Hence context information can be easily incorporated by adjusting the model estimation for the query and the documents.

The risk minimization framework for information retrieval was first proposed by Lafferty and Zhai in [51]. There the risk of returning a single document $\mathbf{d}$ given the query $\mathbf{q}$ is defined as

$$\begin{aligned} R(\mathbf{d};\mathbf{q}) =& R(a=\mathbf{d}|\mathbf{q},\mathcal{C}) \\ =& \sum_{\mathbf{r}\in\{0,1\}} \int_{\theta_Q}\int_{\theta_D} L(\theta_Q,\theta_D,\mathbf{r}) \times p(\theta_Q|\mathbf{q}) \\ & p(\theta_D|\mathbf{d}) p(\mathbf{r}|\theta_Q,\theta_D) d\theta_Q d\theta_D, \end{aligned} \tag{7.1}$$

where $a = \mathbf{d}$ is the action to return the document $\mathbf{d}$ for the query $\mathbf{q}$, $\mathcal{C}$ is the corpus comprising all the documents in the database, $\mathbf{r} \in \{0,1\}$ indicates whether the document $\mathbf{d}$ is relevant to the query $\mathbf{q}$, θ_Q and θ_D are the language models estimated from the query and the document, and L represents the loss function.

Among the many possible loss function definitions, Kullback-Leibler (KL) divergence between the query model and a document model is a well investigated and widely adopted approach [51] and leads to a particularly flexible framework to incorporate additional (e.g. context) information into the retrieval model. Based on the KL divergence loss function and after some approximation of Eqn. (7.1), the following risk function can be obtained:

$$R(\mathbf{d};\mathbf{q}) \propto -\sum_{w_i} p(w_i|\hat{\theta_Q}) \log p(w_i|\hat{\theta_D}) + \xi_q. \tag{7.2}$$

Here, w_i are the words used to represent the query and the documents. Furthermore, ξ_q is a query-dependent constant and can therefore simply be ignored without affecting the ranking result. Finally,

$$\begin{aligned} \hat{\theta_Q} &= \arg\max_{\theta_Q} p(\theta_Q|\mathbf{q}) \\ \hat{\theta_D} &= \arg\max_{\theta_D} p(\theta_D|\mathbf{d}) \end{aligned} \tag{7.3}$$

are the maximum a posteriori estimates of the query and document language models. By assuming a uniform prior, the query and document model can be estimated based on the maximum likelihood principle:

$$\begin{aligned} p_{ml}(w_i|\hat{\theta_Q}) &= \frac{tf(\mathbf{q}, w_i)}{M_q} \\ p_{ml}(w_i|\hat{\theta_D}) &= \frac{tf(\mathbf{d}, w_i)}{M_d}, \end{aligned} \tag{7.4}$$

where $tf(\mathbf{q}, w_i)$ and $tf(\mathbf{d}, w_i)$ are the term frequencies of the word w_i in the query $\mathbf{q}$ and the document $\mathbf{d}$, respectively, and M_q and M_d are the number of words in the query $\mathbf{q}$ and document $\mathbf{d}$, respectively.

For those words which appear in the query and not in the document, the corresponding item $p(w_i|\hat{\theta_Q}) \log p(w_i|\hat{\theta_D})$ in Eqn. (7.2) will have infinite value since $p(w_i|\hat{\theta_D}) = 0$. This will lead to the ranking score $R(\mathbf{d}; \mathbf{q}) = -\infty$. To handle this case, smoothing [138][31] is introduced, the basic idea of which is to assign a prior probability for those words that are "unseen" in documents.

As suggested in the study of language modeling approaches for text and image retrieval [138][31], the Jelinek-Mercer smoothing method performs the best in the image retrieval context as well as for text retrieval with a long query. Since our local query context contains a lot of words, we adopt this method in this chapter. The Jelinek-Mercer smoothing is a linear interpolation of the language model empirically estimated from the documents and the collection model estimated from the whole collection of indexed documents, and can be formulated as

$$p_\lambda(w_i|\hat{\theta_D}) = (1 - \lambda) p_{ml}(w_i|\hat{\theta_D}) + \lambda p(w_i|C), \tag{7.5}$$

where $\lambda \in [0, 1]$ is the trade-off parameter to control the contribution of the smoothing term and $p(w_i|C)$ is the collection language model estimated based on word counts in the entire corpus.

By integrating the smoothed language models Eqn. (7.5) into the risk function defined in Eqn. (7.2) and after some transformation of the risk expression, image documents can be efficiently ranked based on the relevance score computed as follows:

$$\begin{aligned} &S(\mathbf{d}; \mathbf{q}) \stackrel{\text{def}}{=\!=} \\ &\sum_{w_i \in \mathbf{q} \cap \mathbf{d}} p(w_i|\hat{\theta_Q}) \log\left(\frac{(1 - \lambda) p_{ml}(w_i|\hat{\theta_D}) + \lambda p(w_i|C)}{\lambda p(w_i|C)}\right). \end{aligned} \tag{7.6}$$

7.4.2 Context-aware image retrieval model

The incorporation of the local query context into the web image retrieval scenario leads to bi-modal representations of the query and a document, i.e., visual representation of an image and textual representation of its local context. In this respect, we can expand the definition of a query $\mathbf{q}$ as consisting of the query

image $\mathbf{q}^v$, represented by a bag of visual words $\{w_i^v\}$ and the local context $\mathbf{q}^t$ consisting of four fields defined in Section 7.3.1, each of which is represented by a bag of textual words $\{w_i^t\}$. Similarly, each of the images $\mathbf{d}$ from the corpus $\mathcal{C}$ is represented by the bags of visual ($\mathbf{d}^v$) and textual words ($\mathbf{d}^t$).

If the visual and textual words are combined into a unified word space, then the problem of web image retrieval using local and global context can still be formulated using the elegant risk minimization framework from Section 7.4.1. Following this approach, we estimate the expanded query and document models as follows:

$$\begin{aligned} \hat{\theta_Q} &= \arg\max_{\theta_Q} p(\theta_Q|\mathbf{q}^v, \mathbf{q}^t) \\ \hat{\theta_D} &= \arg\max_{\theta_D} p(\theta_D|\mathbf{d}^v, \mathbf{d}^t). \end{aligned} \tag{7.7}$$

Compared to common alternatives for exploiting multi-modal information for web image retrieval, such as reranking [121], which first ranks the images based on one modality and then reranks them by considering the other modality, our adopted approach building on a unified word space not only supports more efficient one-step image retrieval, but also enables the investigation of the expected correlation between visual words and textual words, as will be shown in Section 7.5.2.

In the next section, we elaborate in more detail on how we approach the estimation of the query model using the local and global query context.

7.5 Query Model Estimation using Local and Global Query Context

In Section 7.5.1 we first define the query model incorporating the local query context that unifies the textual context and visual content of the query image into one model and transforms web image search into a structured document retrieval problem. Then, in Section 7.5.2 we define a multi-modal relevance model that integrates the local and global query context and can address the potential limitations facing the local query context.

7.5.1 Query model using local context

Web image retrieval using local query context can be regarded as a structured document retrieval problem [118], where not only the query but also the documents in the database are structured. Here we use the local context not only to enrich the query representation, but also to enrich the document representation. As defined in Section 7.3.1, the local context of image documents consists of four fields: Page title, Alt text, Surrounding text, and Query Association from click-through logs. They are combined with the visual representation to form the five fields in the structured query and document. A straightforward approach

to structured document retrieval is to linearly combine the relevance scores computed on each field separately, which is also referred to as score combination [118][69][42][117][19][131]. Though the approach seems effective, Robertson et al. [81] pointed out that it has several essential drawbacks, such as breaking the non-linearity of term weight and leading to non-robust estimation of collection statistics. Based on this, they proposed to combine the term frequencies (tf) in the *BM25* retrieval model [80]. In this chapter, by extending the tf combination idea to language modeling approach, we propose the model combination method to combine the language models estimated from each field individually:

$$p_l(w_i|\hat{\theta_Q}) = \sum_{k=1}^{K} \alpha_k \times dl_k \times p(w_i|\hat{\theta_Q^k}), \tag{7.8}$$

$$p_l(w_i|\hat{\theta_D}) = \sum_{k=1}^{K} \alpha_k \times dl_k \times p(w_i|\hat{\theta_D^k}), \tag{7.9}$$

where $p(w_i|\hat{\theta_Q})$ and $p(w_i|\hat{\theta_D})$ are the local context based language models for the query image and database image respectively, and k represents the index of the K fields. We set $K = 5$ in this chapter in view of the four context elements we defined in Section 7.3 as well as the visual query representation. $\hat{\theta_Q^k}$ and $\hat{\theta_D^k}$ are the language models for the query and document estimated using the k-th field. α_k is the weight to represent the importance of the k-th field. dl_k is the average document length for the k-th field. It is important to note that w_i can be either a textual word or a visual word.

The incorporation of the average document length is important to balance the different fields. For example, on one of the collected datasets, we found that the average document length of the visual representation is 996 while it is only 2 for Alt text. This means that the average word probability is 1/996 for visual and 1/2 for Alt text. Without scaling by document length, the value of KL divergence computed using Eqn. (7.6) would be dominated by the words appearing in the Alt text. This problem can be addressed by incorporating the average document length to scale the word probability.

The collection language model is estimated in the same way, namely,

$$p_l(w_i|C) = \sum_{k=1}^{K} \alpha_k \times dl_k \times p(w_i|C_k), \tag{7.10}$$

where $p(w_i|C_k)$ is the collection language model estimated from the k-th field.

Since one image may be associated with several web pages, the fields extracted from the web page may be repeatable. Moreover, in the click-through log for each keyword query the images may be clicked by several users, which makes the Query Association field also repeatable. A possible way to deal with the repeatable fields is to weight the sources differently according to the confidence of the source, such

as the PageRank [76] of the web page, and the reliability of the user performing the click. In this chapter, we employ the simplest possible solution that merges the content from different sources (web pages and users) into a single field.

7.5.2 Query model using local and global query context

Although incorporating the local context can effectively improve the web image retrieval performance, it can be problematic in some cases. First, not all images have related web pages or Query Associations. In such cases, it is difficult to obtain the textual description of an image. Second, for those images with local context, the contextual information may be noisy and unable to describe the image.

In this section we explain how we employ the global context to cope with the aforementioned limitations of the local context. Motivated by the relevance-based language model [54], we propose a multi-modal relevance model to incorporate the local and global visual and textual information into the image retrieval model. Different from a relevance-based language model for which the query is a single medium, e.g., text, the multi-modal relevance model can handle a multi-modal representation of queries and documents. Different from the traditional language model, multi-modal relevance model can leverage the knowledge in the corpus to estimate a more comprehensive and robust query model.

The query model should be estimated conditioned by the query image itself as well as the associated textual representation.

$$p(w_i|\{\mathbf{q}^t, \mathbf{q}^v\}) = \frac{p(w_i, \mathbf{q}^t, \mathbf{q}^v)}{p(\mathbf{q}^t, \mathbf{q}^v)}, \tag{7.11}$$

$$p(\mathbf{q}^t, \mathbf{q}^v) = \sum_{w_i} p(w_i, \mathbf{q}^t, \mathbf{q}^v), \tag{7.12}$$

where w_i is either a visual or a textual word. The multi-modal relevance model is now proposed to incorporate the corpus to achieve a more comprehensive and robust query model estimation:

$$\begin{aligned} &p_g(w_i, \mathbf{q}^t, \mathbf{q}^v) = \\ &\sum_{\{\theta^t, \theta^v\} \in \mathcal{M}} p(\theta^t, \theta^v) p(w_i|\theta^t, \theta^v) p(\mathbf{q}^v|\theta^v) p(\mathbf{q}^t|\theta^t), \end{aligned} \tag{7.13}$$

where $\{w_i\} = \{w_i^t\} \cup \{w_i^v\}$ and

$$\begin{aligned} p(w_i^v|\theta^t, \theta^v) &= p(w_i^v|\theta^v), \\ p(w_i^t|\theta^t, \theta^v) &= p(w_i^t|\theta^t). \end{aligned} \tag{7.14}$$

The above equations (Eqn. (7.11–7.14)) estimate the visual query model and contextual query model simultaneously. Here $p(\mathbf{q}^v|\theta^v)$ and $p(\mathbf{q}^t|\theta^t)$ is the visual

and contextual query likelihood, respectively, given the model of a document from the corpus.

The likelihood of the query given the document model in the visual domain can be estimated by assuming that the visual words are independent of each other.

$$p(\mathbf{q}^v|\theta^v) = \prod_{w_j^v \in \mathbf{q}^v} p(w_j^v|\theta^v). \tag{7.15}$$

Since the textual representation is structured with several fields, the likelihood in the textual domain can be estimated as the products of the likelihoods estimated in different fields,

$$p(\mathbf{q}^t|\theta^t) = \prod_k (\prod_{w_j^t \in \mathbf{q}_k^t} p(w_j^t|\theta^t))^{\alpha_k}, \tag{7.16}$$

where α_k is the weight to denote the importance of different fields, similarly to Eqn. (7.9). If the local context $\mathbf{q}^t$ is unavailable, we can simply set $p(\mathbf{q}^t|\theta^t)$ to a small constant for any θ_t and so it can be ignored in the equation.

The model space $\mathcal{M}$ is the space comprising all the document models, which can be regarded as a knowledge base describing the mutual interpretation of images and texts, but also capturing the co-occurrence among the (both visual and textual) words in the corpus. We can approximate $\mathcal{M}$ using all the document models from the images in the corpus and assume that the generation probability $p(\theta^t, \theta^v)$ of image documents is uniform. Then the relevance model defined in Eqn. (7.11–7.14) can be regarded as the weighted sum of the document models from the corpus, where the weights are determined by the likelihood $p(\mathbf{q}^v|\theta^v)p(\mathbf{q}^t|\theta^t)$ of the query given these document models.

Different from the traditional information retrieval where the query typically contains only a small number of keywords, in our approach both the image query and its local context are much more informative and, as such, could also be included in the model space $\mathcal{M}$. Since, however, our experiments indicated that the query likelihood given the query model $p(\mathbf{q}^v|\theta_Q^v)p(\mathbf{q}^t|\theta_Q^t)$ is several magnitudes larger than that given other document models, our proposed multi-modal relevance model will degrade to the model dominated by the local context and the query image. In order to avoid this, but still benefit from the entire available contextual information, we devise the following strategy. We first estimate the multi-modal relevance model using Eqn. (7.11) and Eqn. (7.13) by removing the query model from $\mathcal{M}$ and then combine it with the local context based model defined in Eqn. (7.8). In our experiments, they are linearly combined and the weights are 0.875 and 0.125 for the local and global context based models, respectively.

Compared to the traditional language model, the context-aware query model described above has the following main advantages. First, the power of the underlying multi-modal relevance model is that the query is essentially "reconstructed"

from the document models from the corpus. In this way, it can lead to an alternative query model that has less dependence on the local context, using which potential problems emerging from insufficiently representative query image or imperfect local context could be reduced. This is possible because the noisy textual or visual words are likely to be isolated and seldom appear in the document models simultaneously with other query words. As a consequence, the noisy words will have a relatively low probability in the relevance model. Second, the global context can still estimate a textual query model even if the local context is unavailable. As such, the multi-modal relevance model can be said to perform the multi-modal query expansion, by expanding the textual query words based on the textual, but also on the visual representation of the images in the corpus. In fact, our experiments indicated that even the query model based on the global context only performed surprisingly well. We will elaborate on this in more detail in Section 7.6.

7.5.3 Implementation

A naïve implementation of a multi-modal relevance model will be computationally expensive since it will traverse all the documents in the corpus to compute the summation in Eqn. (7.13). Here we introduce some implementation considerations to make the proposed approach able to retrieve images in real-time.

We can see that the element in the summation in Eqn. (7.13) consists of two components, the generation probability of words given the document and the query likelihood. In other words, the word probability in the relevance model is the weighted sum of word probabilities in each document, weighted by the query likelihood of that document. The relative magnitude of query likelihood can be determined based on the relevance score computed using KL divergence [137]. It can be observed that except the top T results in the ranking list based on KL divergence, the value of the query likelihood given the documents would be low. Hence we can approximate the model space $\mathcal{M}$ using only the T documents with the highest relevance scores computed based on the query image and the local context. Then the computation of Eqn. (7.13) can be completed in real-time. In our experiments, T is set to 100.

The size of the word space, i.e., the number of candidate words for which we need to compute the generation probability in Eqn. (7.13), is another key factor to influence the computational cost. The computational cost of Eqn. (7.13) is linear to the number of candidate words. Moreover, given a relevance model comprising a large number of words, the retrieval time will be very high since in this case nearly all the documents in the database will have overlapping words with the query and so the inverted file index will not be able to speed up the retrieval. Instead of computing the probabilities for all the words, the size of which will be millions for both textual and visual words, in this chapter, we adopt two strategies to limit the word space:

Table 7.1: *Labeling criteria for various relevance levels*

Level	Labeling criteria for an image-image pair
Excellent	Same product with the same brand, the same model
Relevant	Same product with the same brand, different models
Related	Same product with different brands
Fair	Different products with the same brand
Irrelevant	Unrelated images

- The space of visual words is limited to those words appearing in the query image.
- The space of textual words is not expanded unless the local context is unavailable, since, if available, the local context is typically already rich enough. For those queries without the local context, we choose 20 textual words from the local context of the top 10 documents with a large term frequency and small collection probabilities and use them for expansion.

7.6 Experiments

To demonstrate the effectiveness of the proposed context-aware web image retrieval approach, we perform several experiments on two representative web image search datasets. We compare different methods, including textual search, visual search, local context based search, global context based search and local-global context based search. The results reported below provide insights on how to integrate various context categories into real-world web image search applications. Moreover, they demonstrate that the proposed context-aware retrieval model is a promising way to incorporate the contextual information for more reliable content-based web image retrieval.

7.6.1 Dataset

Since there are no publicly available image collections with the associated web pages, we collected an image dataset using a commercial web image search engine. Since the product image search has gained increasing importance and, in addition, is characterized by clear definitions of search relevance, we built our dataset using product images. Twenty product brand names, which are well-known and frequently searched on the web, are selected as queries to crawl up to 1000 images per brand from the web. Then, five images were selected for each product name to be as diverse as possible to represent e.g. different product models or ranking positions in the text-based search result. After completing manual labeling of the collection, we found that for several queries it was difficult to recognize the product model and label relevant images. Hence they were removed from the query

set. Finally the dataset, further referred to as the **Prod60** dataset, was generated that comprises 19069 images, from which 60 images covering twelve product brand names like *BMW, Colgate, Giant, Gillette, Marlboro, Nikon, Heineken, iPod, Macbook, Nike, Whisky, xBox* have been selected as queries.

To simulate the diversity of contexts in real-world applications, i.e. some of images may have high quality context while the contexts of others may be noisy, we expanded the Prod60 dataset as follows. Firstly, for each of the 60 query images, we searched for the duplicates in the web image database again using a commercial image search engine. This resulted in 4632 duplicate images. Then the duplicates images were added to the Prod60 dataset to enrich the queries. We refer to this new dataset further as the **Prod4692** dataset.

For all of these images, we found all the web pages that have links to them and downloaded the web pages for analysis using a private service available in a commercial web image search engine. Then the Page title, Alt text, and surrounding text of the images were extracted. To generate the Query Association data, we used the click-through log from a commercial keyword-based image search engine, and processed it into the format of Query Association for each of the images in our two datasets. Due to the dead link or other reasons, the corresponding web pages could not be downloaded for some of the images. Moreover, the Query Association was unavailable for those images that have never been clicked by any user. This phenomenon made the dataset realistic since in real-world applications one cannot assume that the local context is available for all the images.

Images in the collections were labeled manually by 6 vendors regarding their relevance to each query using five relevance levels. The labeling criterion for each relevance level is given in Table 7.1. For example, if the query image is a car BMW 320i, then all other images of BMW 320i are labeled as *Excellent*, while the images of other BMW cars with different models, such as BMW M6, are labeled as *Relevant*. Other car brands, such as Lamborghini, are all labeled as *Related*. The other images from BMW company, such as BMW logo and BMW bike are labeled as *Fair*. Finally all other images are labeled as *Irrelevant*.

7.6.2 Experimental setup

To investigate the effects of local and global contexts in web image search, we conducted several experiments. The visual search baseline (*Visual*) simply deploys the language modeling approach based on the extracted visual words, inverted file indexing and the Jelinek-Mercer smoothing [31] to retrieve similar images. The text search baseline (*Text*) deploys the textual information in the local context of query images and database images using the retrieval model for structured documents, as shown in Eqn. (7.8) and Eqn. (7.9), without considering the visual features. Since for some query images (11 in the Prod60 dataset) the local context is unavailable, no results could be obtained using text search. For these queries, we used the result from the visual search for the evaluation of text search so that a fair comparison is possible. For investigating the impact of the local context, we not

only studied the approach proposed in this chapter based on model combination (*Loc_m*), but also the method based on score combination (*Loc_s*) serving as a reference for comparison. Finally, we tested two variants of the multi-modal relevance model, the *Global* variant using only the global context and the *GLC* variant in which the local and global context were combined. The parameters for all the methods were set globally and fixed for all queries. We used different smoothing parameters λ in the language models for visual and textual fields, and they are 0.98 and 0.005 respectively.

In addition to the methods mentioned above, we also experimented with the Pseudo Relevance Feedback (PRF) method [121], which was slightly modified to use both the local and global context by means of reranking. The images at the top and the bottom of the ranked results list returned by *Loc_m* were regarded as pseudo-positives and pseudo-negatives, respectively. Then, the Support Vector Machine (SVM) [109] with the RBF (Radial Basis Function) kernel was used to train a visual classifier. Finally, a combination of the prediction scores of SVM and *Loc_m* was used to rank the images.

To evaluate the methods, two well-known measures, Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain [41], were used. MAP is the mean of Average Precisions (AP) computed for each query, which is defined as the area under the non-interpolated precision/recall curve. Since AP can only work with binary relevance judgments, we defined "Irrelevant" and "Related" images as negative while all others were marked as positive. In the results reported below we computed MAP for top N results and we set N to 40. Compared to MAP, NDCG accepts varying relevance levels. For a given query q, NDCG@k is defined as: $\text{NDCG@}k = \frac{1}{Z}\sum_{j=1}^{k}\frac{2^{r(j)}-1}{\log(1+j)}$, where $r(j)$ is the relevance level of the jth document, Z is the normalization coefficient to make the NDCG of a prefect ranking become equal to one, and k is the truncation level.

We extracted the SIFT features using the feature extraction tool provided by Oxford University[2] and used the Robust Approximate Kmeans [56] to cluster the SIFT features into a visual codebook comprising 1M visual words. For the text processing we removed the stop words and performed stemming.

7.6.3 Performance comparison

The performance comparison of the methods described in the previous section on the Prod60 dataset and in terms of NDCG and MAP is shown in Fig. 7.3. The NDCG results indicate that *Text* performs comparably with *Visual* at small truncation levels and outperforms *Visual* at larger truncation levels. The results on the Prod4692 dataset, as shown in Fig. 7.4, further confirm this observation for large truncation levels, while showing that for small truncation levels (less than 14) *Visual* now performs even much better than *Text*. *Visual* can be said to achieve good performance on top results since it can accurately retrieve

[2]http://www.robots.ox.ac.uk/ vgg/research/affine/detectors.html

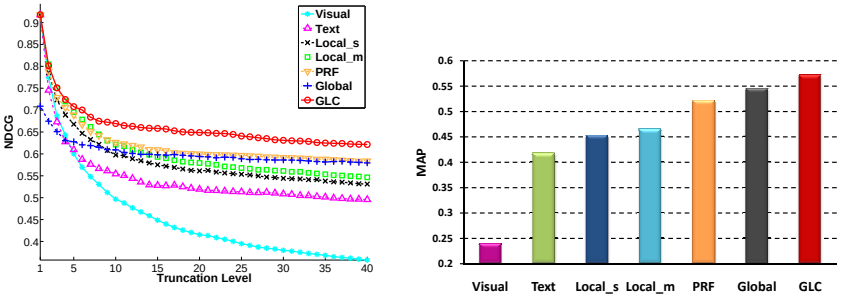

Figure 7.3: *NDCG and MAP comparison over the Prod60 dataset.*

Table 7.2: *Quantitative performance comparison of different algorithms.*

	Prod60						
	Visual	*Text*	*Loc_s*	*Loc_m*	*PRF*	*Global*	*GLC*
NDCG@10	0.497	0.554	0.597	0.621	0.626	0.609	0.669
NDCG@20	0.415	0.519	0.560	0.579	0.600	0.594	0.649
NDCG@40	0.358	0.496	0.531	0.547	0.584	0.579	0.621
MAP	0.240	0.418	0.453	0.465	0.521	0.545	0.572
	Prod4692						
	Visual	*Text*	*Loc_s*	*Loc_m*	*PRF*	*Global*	*GLC*
NDCG@10	0.480	0.465	0.549	0.578	0.568	0.545	0.615
NDCG@20	0.406	0.425	0.499	0.537	0.532	0.532	0.587
NDCG@40	0.349	0.399	0.464	0.516	0.511	0.524	0.566
MAP	0.233	0.318	0.369	0.429	0.430	0.491	0.515

near duplicates and return them on the top. Although *Text* can better convey semantic information for the query image, this information is noisy and not so discriminative to differentiate between different relevance levels.

Since *Visual* is advantageous on retrieving *Excellent* results and *Text* performs better for semantically relevant images, they are expected to complement each other. This has also been confirmed by our results, demonstrating that the local context based methods *Loc_m* and *Loc_s* both significantly outperform *Visual* and *Text*. In addition, we can see that the model combination (*Loc_m*) performs better than the score combination strategy (*Loc_s*), which confirms our hypothesis and Robertson's observations as discussed in Section 7.5.1.

The global context based model (*Global*), without including the local context, performs surprisingly well on both datasets. It performs comparably or slightly

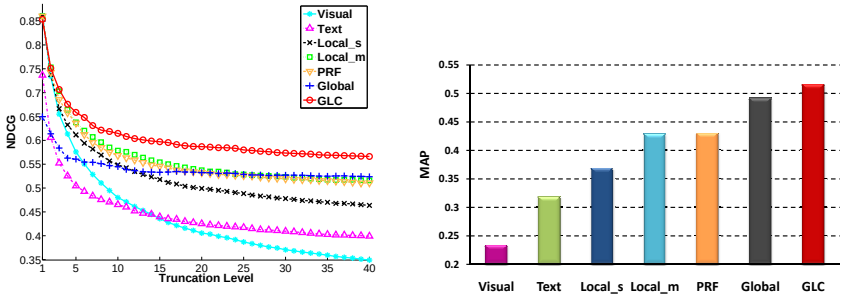

Figure 7.4: *NDCG and MAP comparison over the Prod4692 dataset.*

better than *Loc_m* on NDCG at larger truncation levels, while *Loc_m* performs better on the top results. This indicates that the corpus is a very useful context information source for web image search and that the multi-modal relevance model is indeed able to "reconstruct" the search intention from the corpus. Moreover, we can see that the models biased towards the local and global contexts can be considered complementary to each other. As expected, *GLC* that combines the local and global context can achieve the best performance over all five methods. As shown by the results in Table 7.2, the performance of *GLC* improves by 7.3% compared to *Global*, 13.7% compared to *Loc_m*, 17.0% compared to *Loc_s*, 25.4% compared to *Text*, and 73.8% compared to *Visual* in terms of NDCG@40, and by 4.9%, 23.1%, 26.2%, 36.9%, 138.9% in terms of MAP, respectively. Furthermore, *GLC* outperforms *PRF* by 9.8% in terms of MAP, which demonstrates the effectiveness of the proposed algorithm utilizing the local and global context compared to a well-known alternative method. Moreover, we can see that even *Global*, which uses only the global context, can outperform PRF. This can be explained by the integrated use of the visual and textual features in *Global*.

While *GLC* outperforms *Loc_m* for the truncation levels 10 to 40, it performs comparably with *Loc_m* on the top 10 ranked results. This further confirms that *GLC* leverages the advantage of *Loc_m* on returning relevant images on top results. For those images which can not be certainly determined as relevant by *Loc_m*, incorporating the global context by mining useful information from the corpus further improves their relevance estimation. These conclusions were confirmed by the results obtained on the Prod4692 dataset, as shown in Fig. 7.4. Moreover we can see that on the expanded dataset the local context is noisier than in Prod60. The MAP of *Text* on Prod4692 dataset is only 0.3183, compared to 0.4179 on Prod60. Hence we can say that the proposed method incorporating both the local and global context is generally effective, even though the local context may be noisy.

The performance for each query is presented in Fig. 7.5. There, only the NDCG@40 results are reported because MAP results suggest similar conclusions.

We can observe that *GLC* performs better than *Visual* for 43 queries, and better than *Text* for 47 queries. We can conclude that *GLC* not only significantly improves the overall performance over visual and textual baselines, but also shows superior performance on a large majority of queries.

The reason that on a small amount of queries *GLC* performs worse than the textual or visual baseline can be explained in two ways. First, although the query expansion generates more information to enrich the query, it also brings noise into the retrieval process, especially when the textual or visual search results are extremely poor. Second, when a region in the query image, which does not correspond to users' search intent, finds many duplicates in the corpus, bad visual search result will mask the good textual search result and mislead the *GLC*. In our future work we will address these problems by adaptively adjusting the parameters based on the automatic discovery of the utilities of different categories of contexts.

7.6.4 Analysis

In this subsection, we present a detailed analysis of the impact of various categories of contexts on the retrieval performance. Due to the space limit we will only present the results obtained on the Prod60 dataset. The results on Prod4692 suggest similar conclusions.

Our experimental study of the importance of various fields including Page title, Alt text, Surrounding text, and QA showed that the optimal weights of these fields are 0.01 for Page title, 0.5 for Alt text, 0.09 for Surrounding text and 0.2 for Query Association, respectively. From this we can hypothesize that Alt text is the most important textual field in web image retrieval. While this finding seems contradictory to our intuition that Query Association derived from users' click-through should be the most reliable information source, it can be explained by the fact that Query Association usually represents the general interpretation of the image. In search engines the queries issued by users are most likely to be general words, which we also verified by checking the query log. On the other hand, Alt text, which is provided by the author of the web page, contains a more image-specific information and is therefore more likely to be useful in determining image relevance to the query. A relatively small contribution of the Page title was expected since this information is intended to describe the entire web page. While Surrounding text can mostly be used to interpret the image and contains a large amount of information, it tends to be noisy, which explains its moderate weight in the retrieval model.

To analyze the relative importance of the visual and textual information for the retrieval performance, we fixed the ratio of the weights between different textual fields. The performance variations with the varying visual-versus-textual weight are illustrated in Fig. 7.6. There, the textual weight is the sum of the weights for the four textual fields in Eqn. (7.8) and Eqn. (7.9). From this figure we can see that visual information provides a smaller contribution to the retrieval performance when evaluated using MAP while larger contribution in terms of

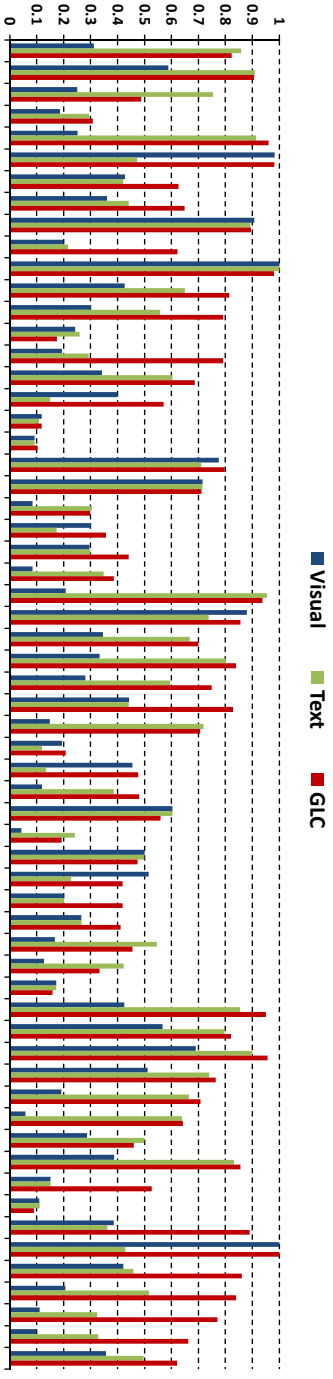

Figure 7.5: *NDCG@40 per query over the Prod60 dataset.*

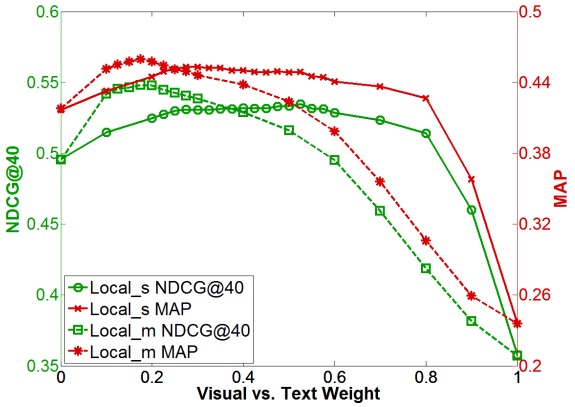

Figure 7.6: *The performance of different combination weights of text and visual models/scores over Prod60. Weight=0 corresponds to purely utilizing the text model/score, while Weight=1 means purely relying on the visual model/score.*

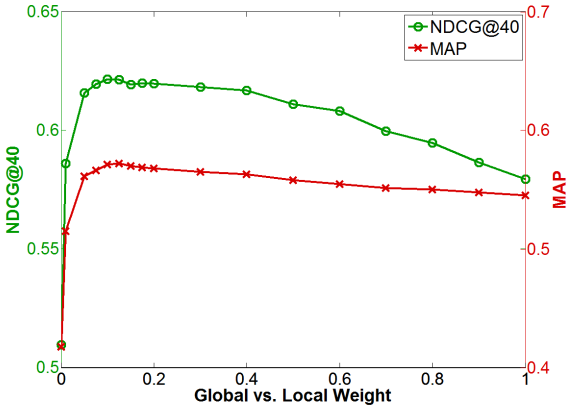

Figure 7.7: *The performance of different combinations of local and global context model over Prod60. Weight=0 means purely utilizing the local context model* Loc_m, *while Weight=1 equals to the global context model* Global.

the NDCG. Since NDCG cares more about the Excellent images than MAP, this confirms the conclusion that visual representation is more suitable for bringing Excellent images to the top of the ranked list.

The effect of the relative weight of global context over local context is shown in Fig. 7.7. Firstly, we can see that, in general, *Global* performs better than *Loc_m*, achieving 13.7% and 30.6% improvement in terms of NDCG@40 and MAP, respectively. However, by combining them together with suitable weights, we can achieve a significant performance improvement. The optimal weight of the global

context over the local context can be set to be between 0.1 and 0.2 and the performance is not sensitive to the weight if it is larger than 0.1.

7.7 Conclusion

With the objective of improving content-based web image retrieval based on the QBE paradigm, we studied in this chapter the possibilities to enrich the query image using various types of contextual information and to embed this information effectively into a context-aware image retrieval model that is based on the language modeling theory. The contextual information we studied includes the associated web page, the click-through log and the corpus. The experimental results on a collected web image dataset demonstrated the utility of the contextual information sources we considered in this chapter and the effectiveness of the proposed context-aware image retrieval model compared to the text-only or visual-only web image retrieval.

Chapter 8

Context in Image Retrieval: Reflections and Recommendations

The word "context" is from Latin *contextus*, the past participle of *contexere*, where *con* stands for "together" and *texere* stands for "to weave". This origin of the word then also provides a recipe for how to effectively leverage context in example-based image search. As indicated by *con*, we first need to discover useful additional sources of information about the query object and then we need to develop an effective model to weave the information derived from the context and from the analysis of the query in order to improve the image retrieval performance.

According to Melucci [70], the first of the two steps mentioned above corresponds to identifying the *contextual variables*, which are the other *observable* in the search process in addition to the query itself. The reason of introducing context in the search process is that it is difficult to infer the users information need accurately from the query alone, especially if the query is semantically as complex as an image. Among the context signals potentially useful for the example-based image retrieval, we focus in this thesis on discovering and deploying two categories of these signals that can be referred to as the query-internal and query-external context. In addition to demonstrating the impact of different course of contextual information on the performance of web image search, another important insight provided in Part II of this thesis is how language modeling can be deployed to incorporate context into conventional retrieval models.

The query-internal context was first investigated in Chapter 5 and was searched for in the image content surrounding the target object (the region of interest). While the usability of this information for improving object-based image retrieval has been recognized before, the reports found in the literature on embedding it into the image retrieval process indicated that there is still significant room for improvement. Our proposed method show how this improvement can be achieved,

and then especially when the query object is small or occluded. While Chapter 5 considers the query that consists of a single image only, we show in Chapter 6 how further improvement can be achieved by expanding the query to an entire video captured about the target object. In this video-based image retrieval approach, the query representation is enriched using the information extracted from the context of all frames of the video.

The query-external context includes the contextual information extracted from different channels than the query itself. The local and global query context we introduced in Chapter 7 belong to this category, and were shown to be helpful to improve semantic image retrieval. While it is easy for a human to interpret an image at the semantic level, inferring this semantics automatically by a computer from image pixels only is challenging and in many cases even impossible. The method presented in Chapter 7 showed how textual information associated with the image, ranging from the texts on the hosting web pages to the textual information collected from the broader web can be discovered, organized and deployed to help infer the semantic of the image query.

While Part II of the thesis proposes a number of innovative approaches on how to deploy context for improved web image search, we conjecture that still many challenges remain open. This leads to a number of promising research topics that we briefly mention here. First, the context acquisition as reported in this thesis is largely passive, in the sense that it is extracted from the data that are available after the users interaction with data, e.g. after the data (image) was captured. To further improve the retrieval performance, it would be beneficial to also consider and capture in some way the actual use context directly at the interaction time. Video-based image retrieval did a preliminary attempt in this direction. However, we hope to see a larger progress there, for instance following the approach of Kofler et al. [48] where query quality is assessed from the log of the search session – the idea that could also be deployed in a more general scope to infer and model the search context of the user. Second, mechanisms are needed for automatically assessing the utility of different available contextual information channels and for recommending the way and extent to which a given contextual channel should be taken into account in a given image search case. This would enable the development of fully adaptive context-aware retrieval models that optimally make use of the available resources to learn the users information need and act accordingly.

Bibliography

[1] http://images.bing.com/.

[2] http://images.google.com/.

[3] http://images.yahoo.com/.

[4] http://www.flickr.com/.

[5] http://www.google.com/mobile/goggles/.

[6] http://www.robots.ox.ac.uk/ vgg/data/oxbuildings.

[7] http://www.tineye.com.

[8] Trec-10 proceedings appendix on common evaluation measures. *http://trec.nist.gov/pubs/trec10/appendices/measures.pdf*.

[9] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, B. Croft, S. Dumais, N. Fuhr, D. Harman, D. J. Harper, D. Hiemstra, T. Hofmann, E. Hovy, W. Kraaij, J. Lafferty, V. Lavrenko, D. Lewis, L. Liddy, R. Manmatha, A. McCallum, J. Ponte, J. Prager, D. Radev, P. Resnik, S. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, and C. Zhai. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst, september 2002. *SIGIR Forum*, 37(1):31–47, 2003.

[10] X. Bai and G. Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV '07: Proceedings of the 11th IEEE International Conference on Computer Vision.*, pages 1–8, 2007.

[11] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006.

[12] N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.

[13] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[14] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, Cambridge, MA, 2008.

[15] J. Callan and N. Belkin. Context-based information access. *Report of the Discussion Group on Context-Based Information Access of the Workshop on "Information Retrieval and Databases: Synergies and Syntheses"*, 2003.

[16] L. Cao, J. Luo, H. Kautz, and T. Huang. Image annotation within the context of personal photo collections using hierarchical event and scene models. *IEEE Transactions on Multimedia*, 11(2):208–219, Feb. 2009.

[17] Z. Cao and T.-Y. Liu. Learning to rank: From pairwise approach to listwise approach. In *ICML*, 2007.

[18] S.-F. Chang, J. He, Y.-G. Jiang, E. El Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. In *NIST TRECVID Workshop*, Gaithersburg, MD, November 2008.

[19] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. In *NIST TRECVID Workshop*, Gaithersburg, MD, November 2006.

[20] O. Chapelle, Q. V. Le, and A. J. Smola. Large margin optimization of ranking measures. In *NIPS Workshop: Machine Learning for Web Search*, 2007.

[21] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *CVPR*, 2007.

[22] J. Cui, F. Wen, and X. Tang. Real time google and live image search re-ranking. In *ACM Multimedia*, 2008.

[23] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[24] M. Davis, S. King, N. Good, and R. Sarvas. From context to content: Leveraging context to infer media metadata. In *Proceeding of the 12th ACM international conference on Multimedia*, MM '04, pages 188–195. ACM Press, 2004.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR '09: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248 –255, jun. 2009.

[26] Z. Dou, R. Song, J.-R. Wen, and X. Yuan. Evaluating the effectiveness of personalized web search. *IEEE Transactions on Knowledge and Data Engineering*, 21:1178–1190, 2008.

[27] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*. IEEE Computer Society, 2005.

[28] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for Google images. In *ECCV*, 2004.

[29] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *CVPR*, 2008.

[30] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding (CVIU)*, 114:712–722, 2010.

[31] B. Geng, L. Yang, and C. Xu. A study of language model for image retrieval. In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pages 158–163, Washington, DC, USA, 2009. IEEE Computer Society.

[32] A. Hanjalic. New grand challenge for multimedia information retrieval: Bridging the utility gap. *International Journal of Multimedia Information Retrieval*, Sep. 2012.

[33] A. Hanjalic, C. Kofler, and M. A. Larson. Intent and its discontents: The user at the wheel of the online video search engine. In *ACM Multimedia*, 2012.

[34] A. Hauptmann, R. Yan, and W.-H. Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *CIVR*, 2007.

[35] S. Heymann, K. Muller, A. Smolic, B. Frohlich, and T. Wiegand. SIFT implementation and optimization for general-purpose GPU. In *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2007.

[36] W. Hsu, L. Kennedy, and S. Chang. Reranking methods for visual search. *Multimedia, IEEE*, 14(3):14–22, 2007.

[37] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *ACM Multimedia*, 2006.

[38] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *ACM Multimedia*, 2007.

[39] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[40] V. Jain and M. Varma. Learning to re-rank: query-dependent image reranking using click data. In *WWW*, 2011.

[41] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR*, 2000.

[42] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. A unified framework for image retrieval using keyword and visual features. *IEEE Transactions on Image Processing*, 2005.

[43] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. on PAMI*, 30(11):1877–1890, 2008.

[44] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

[45] T. Joachims. Training linear svms in linear time. In *KDD*, 2006.

[46] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30, 1938.

[47] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 631–640, New York, NY, USA, 2007. ACM.

[48] C. Kofler, L. Yang, M. Larson, T. Mei, A. Hanjalic, and S. Li. When video search goes wrong: predicting query failure using search engine logs and visual search results. In *ACM Multimedia*, pages 319–328, 2012.

[49] F. Korn and S. Muthukrishnan. Influence sets based on reverse nearest neighbor queries. In *SIGMOD*, 2000.

[50] J. Krapac, M. Allan, J. Verbeek, and F. Jurie. Improving web image search results using query-relative classiffiers. In *CVPR*, 2010.

[51] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119, New York, NY, USA, 2001. ACM.

[52] P. Laskov, C. Gehl, S. Krüger, and K.-R. Müller. Incremental support vector learning: Analysis, implementation and applications. *J. Mach. Learn. Res.*, 7:1909–1936, December 2006.

[53] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 175–182, New York, NY, USA, 2002. ACM.

[54] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, New York, NY, USA, 2001. ACM.

[55] S. Lawrence. Context in web search. *IEEE Data Engineering Bulletin*, 23(3):25–32, 2000.

[56] D. Li, L. Yang, X.-S. Hua, and H.-J. Zhang. Large-scale robust visual codebook construction. In *ACM Multimedia*, 2010.

[57] L.-J. Li and L. Fei-Fei. OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning. *Int. J. Comput. Vision*, 2009.

[58] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 351–360, Madrid, Spain, 2009. ACM.

[59] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3:225–331, March 2009.

[60] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *ICME*, 2008.

[61] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2), 2004.

[62] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proc. of the 1981 DARPA Imaging Understanding Workshop*, 1981.

[63] J. Luo, A. Hanjalic, Q. Tian, and A. Jaimes. Integration of context and content for multimedia management: An introduction to the special issue. *Multimedia, IEEE Transactions on*, 11(2):193 – 195, January 2009.

[64] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *MULTIMEDIA '03: Proceedings of the 11th ACM international conference on Multimedia*, pages 374–381. ACM, 2003.

[65] D. K. Mahajan and M. Slaney. Image classification using the web graph. In *ACM Multimedia*, pages 991–994, 2010.

[66] A. Makadia. Feature tracking for wide-baseline image retrieval. In *ECCV*, 2010.

[67] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008.

[68] T. Mei, X.-S. Hua, and S. Li. Contextual in-image advertising. In *MULTIMEDIA '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 439–448, New York, NY, USA, 2008. ACM.

[69] T. Mei, Z.-J. Zha, Y. Liu, M. Wang, G.-J. Qi, X. Tian, J. Wang, L. Yang, and X.-S. Hua. MSRA at TRECVID 2008: High-Level Feature Extraction and Automatic Search. In *TRECVID*, 2008.

[70] M. Melucci. Contextual search: A computational framework. *Foundations and Trends in Information Retrieval*, 6(4-5):257–405, 2012.

[71] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[72] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, 2009.

[73] A. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *ACM Multimedia*, 2007.

[74] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[75] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, December 2007.

[76] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[77] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[78] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning*. MIT Press, 1999.

[79] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV '07: Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1 –8, oct. 2007.

[80] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389, April 2009.

[81] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49, New York, NY, USA, 2004. ACM.

[82] S. E. Robertson and D. A. Hull. The trec-9 filtering track final report. In *TREC*, 2000.

[83] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, August 2004.

[84] S. Rudinac, M. Larson, and A. Hanjalic. Exploiting noisy visual concept detection to improve spoken content based video retrieval. In *ACM Multimedia*, 2010.

[85] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 2002.

[86] F. Scholer and H. E. Williams. Query association for effective retrieval. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 324–331, New York, NY, USA, 2002. ACM.

[87] F. Scholer, H. E. Williams, and A. Turpin. Query association surrogates for web search: Research articles. *Journal of the American Society for Information Science and Technology*, 55(7):637–650, 2004.

[88] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[89] L. Shang, L. Yang, F. Wang, K.-P. Chan, and X.-S. Hua. Real-time large scale near-duplicate video retrieval. In *ACM Multimedia*, 2010.

[90] H. Shen, B. Ooi, and K. Tan. Giving meanings to www images. In *MULTIMEDIA '00: Proceedings of the 8th ACM international conference on Multimedia*, pages 39–47, 2000.

[91] P. Sinha and R. Jain. Semantics in digital photos: A contextual analysis. In *2008 IEEE International Conference on Semantic Computing*, pages 58 –65, aug. 2008.

[92] J. Sivic, F. Schaffalitzky, and A. Zisserman. Object level grouping for video shots. 2004.

[93] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[94] M. Slaney. Web-scale multimedia analysis: Does content matter? *MultiMedia, IEEE*, 18(2):12 –15, Feb. 2011.

[95] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[96] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[97] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia*, 2006.

[98] D. Stavens and S. Thrun. Unsupervised learning of invariant features using video. In *CVPR*, 2010.

[99] Thorsten Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[100] X. Tian, Y. Lu, L. Yang, and Q. Tian. Learning to judge image search results. In *ACM Multimedia*, 2011.

[101] X. Tian, D. Tao, X.-S. Hua, and X. Wu. Active reranking for web image search. *IEEE Trans. Img. Proc.*, 19:805–820, March 2010.

[102] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *ACM Multimedia*, 2008.

[103] X. Tian, L. Yang, X. Wu, and X.-S. Hua. Visual reranking with local learning consistency. In *MMM*, 2010.

[104] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 273, Washington, DC, USA, 2003. IEEE Computer Society.

[105] J. K. Tsotsos, S. M. Culhane, W. Y. K. Winky, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, October 1995.

[106] P. Turcot and D. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV Workshop (WS-LAVD)*, 2009.

[107] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *WWW*, 2009.

[108] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *WWW*, 2009.

[109] V. N. Vapnik. Statistical learning theory. *Wiley*, 1998.

[110] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org, 2008.

[111] D. Wagner, D. Schmalstieg, and H. Bischof. Multiple target detection and tracking with guaranteed framerates on mobile phones. In *ISMAR*, 2009.

[112] J. Wang and M. F. Cohen. *Image and video matting: a survey*, volume 3. Now Publishers Inc., Hanover, MA, USA, 2007.

[113] L. Wang, L. Yang, and X. Tian. Query aware visual similarity propagation for image search reranking. In *ACM Multimedia*, 2009.

[114] M. Wang, L. Yang, and X.-S. Hua. MSRA-MM: Bridging research and industrial societies for multimedia information retrieval. Technical report, In Microsoft Technical Report, 2009.

[115] X.-J. Wang, W.-Y. Ma, G.-R. Xue, and X. Li. Multi-model similarity propagation and its application for web image retrieval. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 944–951, New York, NY, USA, 2004. ACM.

[116] X.-J. Wang, L. Zhang, X. Li, and W. Y. Ma. Annotating images by mining image search results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1919–1932, 2008.

[117] T. Westerveld, A. P. D. Vries, A. van Ballegooij, F. de Jong, and D. Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *EURASIP Journal on Applied Signal Processing*, 2003:186–198, 2003.

[118] R. Wilkinson. Effective retrieval of structured documents. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 311–317, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[119] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to tag. In *WWW*, pages 361–370, 2009.

[120] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 118–126, New York, NY, USA, 2004. ACM.

[121] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *CIVR*, 2003.

[122] L. Yang, Y. Cai, A. Hanjalic, X.-S. Hua, and S. Li. Searching for images by video. *International Journal of Multimedia Information Retrieval*, pages 1–13, 2012.

[123] L. Yang, B. Geng, Y. Cai, A. Hanjalic, and X.-S. Hua. Object retrieval using visual query context. *IEEE Transactions on Multimedia*, 13(6):1295–1307, 2011.

[124] L. Yang, B. Geng, A. Hanjalic, and X.-S. Hua. Contextual image retrieval model. In *CIVR*, 2010.

[125] L. Yang, B. Geng, A. Hanjalic, and X.-S. Hua. A unified context model for web image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 8(3):28, 2012.

[126] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *ACM Multimedia*, pages 183–192, 2010.

[127] L. Yang and A. Hanjalic. Learning from search engine and human supervision for web image search. In *ACM Multimedia*, 2011.

[128] L. Yang and A. Hanjalic. Prototype-based image search reranking. *IEEE Transactions on Multimedia*, 14(3-2):871–882, 2012.

[129] L. Yang and A. Hanjalic. Learning to rerank web images. *IEEE Multimedia Magazine*, To Appear.

[130] X. Yang, S. Pang, and K. Cheng. Mobile image search with multimodal context-aware queries. In *International Workshop on Mobile Vision (In conjunction with CVPR 2010)*, 2010.

[131] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *MULTIMEDIA '09: Proceedings of the seventeen ACM international conference on Multimedia*, pages 175–184, New York, NY, USA, 2009. ACM.

[132] Y.-H. Yang and W. Hsu. Video search reranking via online ordinal reranking. In *ICME*, 2008.

[133] Y. H. Yang, P. T. Wu, C. W. Lee, K. H. Lin, W. H. Hsu, and H. H. Chen. Contextseer: context search and recommendation at query time for shared consumer photos. In *MULTIMEDIA '08: Proceeding of the 16th ACM international conference on Multimedia*, 2008.

[134] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *ICML*, 2008.

[135] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *MULTIMEDIA '09: Proceedings of the 17th ACM international conference on Multimedia*, pages 15–24, New York, NY, USA, 2009. ACM.

[136] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, and X.-S. Hua. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 6:13:1–13:19, August 2010.

[137] C. Zhai. *Statistical Language Models for Information Retrieval*. Morgan & Claypool, 2009.

[138] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

[139] H.-J. Zhang. Multimedia content analysis and search: New perspectives and approaches. In *"www.acmmm09.org/ACM MM09 Keynote.pdf"*, 2009.

[140] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieva. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 846–851, Washington, DC, USA, 2005. IEEE Computer Society.

[141] R. Zhao and W. I. Grosky. Narrowing the semantic gap - improved text-based web document retrieval using visual features. *IEEE Transactions on Multimedia*, 4(2):189–200, August 2002.

[142] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst.*, 8(6):536–544, 2003.

Summary

While retrieval models are of fundamental importance in Information retrieval, they have not been studied well for the cases when the retrieval concerns *multimedia* data collections. In this thesis we focused on developing advanced retrieval models to help improve the retrieval of images from the Web. Because a typical Web image retrieval system usually supports both keyword queries and example image queries, the thesis is naturally divided into two parts to address the retrieval using both query types.

The first part of the thesis addresses the keyword-based image search, and then specifically the challenge of *image search reranking* in the Web context. The basic principle of reranking is that the image search results list acquired using a textual query is refined using the information extracted from the visual content of the images in the list. The reranking approaches proposed in the past were largely unsupervised. However, the reranking criteria deployed there to determine how visual features of the images are deployed for refining the results list were largely heuristic and therefore insufficiently reliable in a general case. To improve the reliability, in Chapter 2 we introduced the idea of *supervised reranking*, where a human supervision step is embedded in devising the reranking model using eleven carefully designed reranking features. In Chapter 3, this idea is generalized using *prototype-based reranking* techniques. We constructed the so-called *prototypes* from the initial search result and then proposed three ways of building *meta-rerankers* from these prototypes, which are then combined into the final reranking model in a supervised fashion. This part of the thesis is concluded by Chapter 4, which presents a systematic review of the reranking approaches and which identifies the remaining challenges for developing and deploying the reranking technology in real-world image search engines.

In part II, we address the query-by-example image search scenario and focus on discovering and utilizing various contextual signals to help improve the retrieval accuracy. Chapter 5 is devoted to object-based image search, where the visual scene context surrounding the query object (object of interest) is deployed to help find more images of that object. To that end, we developed a contextual

object retrieval model effectively incorporating the visual scene context. In order to acquire even richer contextual information for object-based image retrieval, in Chapter 6 we extend the example-based image search concept into *video-based image retrieval* (VBIR), which allows users to submit not a single image, but a video clip about the query object. Since the video clip shows the target object under varying capture conditions, rich information can be extracted from this clip, through which the search engine can understand the query object better and fine-tune the retrieval model accordingly. In Chapter 7 we generalize our approach of context-aware image retrieval and present a *semantic image retrieval model* that combines both the *local* and *global* context together to better understand images semantics. Here the local and global contexts are mined, respectively, from the Web pages associated with images and the click-through data, which can be regarded as a knowledge base of the search engine. Chapter 8 concludes this part of the thesis by a brief summary and recommendations for future research.

Samenvatting

Hoewel retrievalmodellen van fundamenteel belang zijn in het vakgebied van information retrieval, zijn ze nog niet goed bestudeerd in de gevallen wanneer het gaat om het zoeken in *multimedia*collecties. In dit proefschrift richten wij ons op het ontwikkelen van geavanceerde retrievalmodellen om het vinden van afbeeldingen op het web te helpen verbeteren. Gezien een typisch zoeksysteem voor afbeeldingen op het web doorgaans zowel zoekopdrachten in de vorm van trefwoorden als zoekopdrachten op basis van een voorbeeldafbeelding ondersteunt, is dit proefschrift dan ook opgedeeld in twee delen om beide manieren van zoeken te behandelen.

Het eerste deel van het proefschrift behandelt het vinden van afbeeldingen op basis van trefwoorden en dan in het bijzonder de uitdaging van *het herrangschikken van afbeeldingen* in de context van het web. Het basisprincipe van herrangschikken is, dat de resultatenlijst van afbeeldingen die verkregen is met een tekstuele zoekopdracht, wordt verfijnd met behulp van de informatie gextraheerd uit de visuele inhoud van de afbeeldingen uit die lijst. De manieren om te herrangschikken die in het verleden werden voorgesteld waren voornamelijk ongesuperviseerd. De criteria die voor het herrangschikken werden gebruikt voor het bepalen hoe visuele afbeeldingskenmerken werden ingezet voor het verfijnen van de resultatenlijst, waren echter grotendeels heuristisch van aard. Derhalve waren zij onvoldoende betrouwbaar voor een algemeen geval. Om de betrouwbaarheid te verbeteren hebben wij in hoofdstuk 2 het idee van *gesuperviseerd herrangschikken* gentroduceerd. Hierbij is een menselijke supervisiestap ingebouwd in het ontwikkelen van het model gebruikmakende van elf met zorg ontworpen herrangschikkenmerken. Dit idee wordt algemener gemaakt in hoofdstuk 3 door gebruik te maken van *prototype-gebaseerde herrangschiktechnieken*. Wij construeerden uit de initile zoekresultaten de zogenaamde *prototypes* en stelden vervolgens drie manieren voor voor het opbouwen van *meta-herrangschikkers* uit deze prototypes, die dan op gesuperviseerde wijze worden gecombineerd tot het uiteindelijke herrangschikmodel. Het eerste deel van het proefschrift wordt afgesloten door hoofdstuk 4 dat een systematisch overzicht van herrangschikmethoden

presenteert en dat de resterende uitdagingen vaststelt voor het ontwikkelen en het inzetten van herrangschiktechnologie voor afbeeldingszoekmachines in de praktijk.

In het tweede deel van het proefschrift behandelden we het voorbeeldafbeelding-als-zoekopdracht scenario en concentreren wij ons op het ontdekken en het gebruikmaken van verscheidende contextuele signalen om de accuratesse van het zoeken te helpen verbeteren. Hoofdstuk 5 is toegewijd aan object-gebaseerd zoeken naar afbeeldingen. De visuele context dat het object in de zoekopdracht (het voorwerp waarin de gebruiker is genteresseerd) omringt, wordt hierbij ingezet om te helpen bij het vinden van meer afbeeldingen van dat object. Daartoe ontwikkelden wij een contextueel objectretrievalmodel dat de context van de visuele scne effectief omvat. Om nog rijkere contextuele informatie te verkrijgen voor object-gebaseerd zoeken naar afbeeldingen breiden we in hoofdstuk 6 het concept van voorbeeld-gebaseerd zoeken naar afbeeldingen uit tot *video-gebaseerd image retrieval* (VBIR), waarin gebruikers niet een enkele afbeelding, maar een geheel videofragment over het te vinden object kunnen opgeven als zoekopdracht. Aangezien het videofragment het doelobject onder verschillende opnameomstandigheden laat zien, kan er rijke informatie uit dit fragment worden onttrokken. De zoekmachine kan deze informatie gebruiken om het object in de zoekopdracht beter te begrijpen en het retrievalmodel dienovereenkomstig af te stemmen. In hoofdstuk 7 generaliseren wij onze aanpak van contextbewust zoeken naar afbeeldingen en presenteren een *semantisch retrievalmodel voor afbeeldingen* dat zowel de *lokale* als *globale* context combineert om beter de semantiek van de afbeeldingen te begrijpen. De lokale en globale contexten worden hier respectievelijk ontgonnen van webpaginas geassocieerd met afbeeldingen en van de doorklikdata, wat kan worden gezien als een kennisdomein van de zoekmachine. Hoofdstuk 8 sluit dit deel van het proefschrift af met een beknopte samenvatting en aanbevelingen voor verder onderzoek.

Acknowledgements

Pursuing the PhD degree is a big challenge in life. This challenge is very difficult to accomplish without any help. I was very fortunate to have had the opportunity to work with numerous talented people and to learn from them in the process.

First of all, I would like to thank my thesis advisor, Alan Hanjalic, for all the discussions, suggestions, and guidance about my PhD thesis. Although I have published several papers before enrolling in the PhD program, I still learned a lot from Alan about the skills of conducting solid scientific research and writing high-impact scientific papers. Alans guidance significantly helped my growth towards a well-established researcher, for which I am very grateful.

Second, I would like to thank my thesis co-advisor, Inald Lagendijk, for all the suggestions he gave me about my PhD thesis over the past four years. In particular, I still remember my initial discussion with Inald during my first visit to Delft in 2009, during which I learned how important it is to think critically and rigorously about a research approach and how to deploy this way of thinking to make that approach as solid as possible.

Third, I would like to thank my former managers, Xian-Sheng Hua and Shipeng Li. Xian-Sheng was my first manager at the Microsoft Research Asia (MSRA), where I started my research career. Being a fresh university graduate, I learned a lot from Xian-Sheng about how to become a good researcher. I am also grateful to Xian-Sheng for creating the opportunity for me to pursue my PhD degree in parallel with my regular work at MSRA and for encouraging and supporting me during my PhD thesis research. From Shipeng as our group manager I learned a lot about strategic issues concerning scientific research, and in particular how to think big and strive for long-term scientific impact.

Fourth, I would like to thank my interns at MSRA, Bo Geng and Yang Cai. We had excellent discussions and co-authored several scientific papers together. They also helped me in setting up and running a portion of the experiments reported in the thesis. I am very grateful for their help. I would also like to thank Raynor Vliegendhart, who helped translate the Summary into Dutch.

Last but not least, I would like to thank all the people whom I worked with and

who helped me during the past four years. I specifically thank my wife Ying Liu, whose support is invaluable for me and who makes everything I do worthwhile.

Curriculum Vitae

Linjun Yang was born in Anhui, China on June 16th, 1980. He obtained his B.S degree in Electronics Engineering from East China Normal University, Shanghai, China in 2001 and M.S. degree in Computer Science from Fudan University, Shanghai, China in 2006. From 2006 to 2012, he was with Microsoft Research Asia (MSRA), Beijing, China, first as an assistant researcher and then as an associate researcher in the Media Computing group. In 2012, he joined Microsoft Bing as a Research Software Development Engineer. In 2009, in parallel with his work at Microsoft, he became a part-time PhD student at the Delft University of Technology, The Netherlands.

His current research interests are in the broad area of multimedia information retrieval, with focus on multimedia search ranking and large-scale Web multimedia mining. His research resulted in 13 journal papers, more than 40 conference papers, and 1 book chapter. He received the Best Paper Award from the ACM Multimedia 2009 conference and the Best Student Paper Award from the ACM Conference on Information and Knowledge Management (CIKM) 2009.

Linjun Yang is a member of ACM and IEEE. He served as a Program Committee member or reviewer for many international conferences, like ACM Multimedia, SIGIR, ICMR, CIVR and ICIP, and as a reviewer for scientific journals including IEEE Transactions on Multimedia and IEEE Transactions on Circuits and Systems for Video Technology.