Advancing the Relevance Criteria for Video Search and Visual Summarization

Stevan Rudinac

Advancing the Relevance Criteria for Video Search and Visual Summarization

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 6 mei 2013 om 10:00 uur
door

Stevan RUDINAC

diplomirani inženjer elektrotehnike
Universiteit van Belgrado, Servië
geboren te Foča, Bosnië en Herzegovina, Joegoslavië

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. B. Merialdo,	EURECOM, France
Prof. dr. F.G.B. De Natale,	University of Trento, Italy
Prof. dr. ir. G.J.P.M. Houben,	Technische Universiteit Delft
Prof. dr. P.J. Stappers,	Technische Universiteit Delft
Dr. C.G.M. Snoek,	Universiteit van Amsterdam

Dr. M.A. Larson heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

The research reported in this thesis has received funding from the European Commission's 7th Framework Programme (FP7) under grant agreement no. 216444 (NoE PetaMedia).

This work was carried out in the ASCI graduate school.
ASCI dissertation series number 277.

ISBN 978-94-6186-147-4

Cover design by Jelena Popadić
Printed by CPI Wöhrmann Print Service, Zutphen

Contents

Chapter 1

Introduction

During the information age, as the period since 1970's is frequently called, the amount of information stored in personal and public collections exceeded all expectations. Increasing sophistication and affordability of personal computers, storage media and content capturing devices, as well as the birth of the global communication network, the Internet, have significantly simplified capturing, distribution and consumption of digital content (e.g., text, images, video, music and a combination of these - multimedia), effectively transforming content consumers into the active content producers and publishers. In addition, much of the information previously stored in written, printed or analog form has been digitized and made publicly accessible via Internet. While the written and printed books were for centuries almost exclusive means of documenting and communicating the knowledge, the announcement made in March 2012 that the 2010 edition of Encyclopædia Britannica [1] will be the last printed edition, serves as yet another confirmation of a major paradigm shift in content publishing and consumption.

Besides numerous advantages, the increase in content production and availability has created significant challenges as well, one of the most important being to enable finding the relevant content that satisfies a user's information need. While pursuing this challenge in its entirety generally requires a multi-disciplinary effort, its technological dimension inspired the birth of *multimedia information retrieval* (MIR), a rapidly expanding research direction in computer science. The aim of MIR is developing the (semi-) automated mechanisms to facilitate information finding in multi-

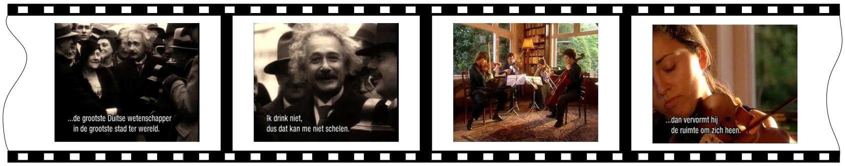

...de grootste Duitse wetenschapper
in de grootste stad ter wereld.
Ik drink niet,
dus dat kan me niet schelen.
...dan vervormt hij
de ruimte om zich heen.

Figure 1.1: Illustration of a documentary video from the S&V archive, consisting of visual, auditory (spoken content and music) and textual (overlay text) modalities.

media collections to the highest possible extent.

This thesis presents the results of our research for which we believe to significantly improve the state-of-the-art in two aspects of multimedia information retrieval, namely, *video search* and *visual summarization*.

1.1 Multimedia and social media

In this section we first go deeper into the content types dealt with in the thesis, namely, *multimedia* and *social media*. The term multimedia refers to a type of content, consisting of multiple content forms (media), such as e.g., text, image, audio and video. Each medium carries information of a particular type, which is communicated to a human via a particular communication channel, commonly called modality [14]. Video is an example of a truly multimodal content form, as it usually consists of visual, auditory (e.g., music and spoken content) and textual (e.g., overlay text and subtitles) modalities. An example documentary video from the Netherlands Institute for Sound and Vision (S&V)[1], consisting of those modalities is illustrated in Fig. 1.1. As combining multiple content forms leads to richer information sources, multimedia quickly became a dominant type of content. Articles in online encyclopedia such as e.g., Wikipedia [2] and Encyclopædia Britannica [1] frequently combine text, images, videos and music for a more complete and appealing presentation.

When searching for solutions for facilitating access to multimedia, it is important to note the difference between *data* (content) and *metadata* (information about the content). An image or a video in a digital collection is frequently associated with various types of metadata, such as e.g., those automatically generated by the content capturing device (e.g., content production date, format and geo-location) and manually generated

[1] www.beeldengeluid.nl

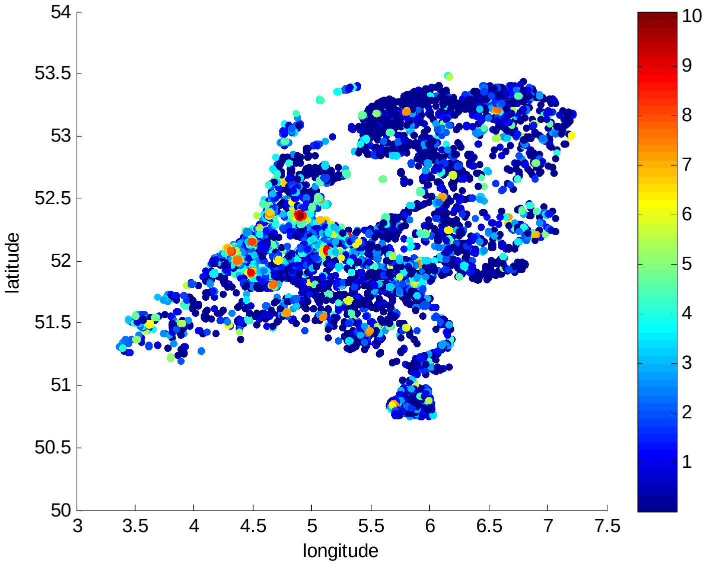
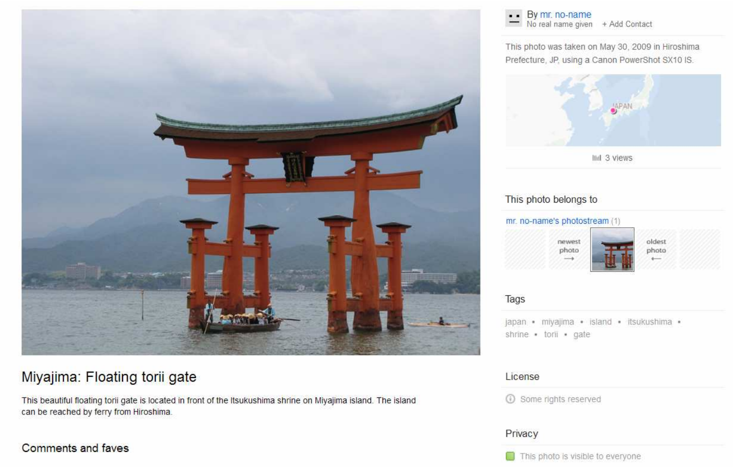


Figure 1.2: Illustration showing the number of geo-referenced Flickr images (in logarithmic scale) captured within 1km from each monument in The Netherlands. Only images available under a Creative Commons (CC) license are taken into account.

textual annotations (e.g., title, description and keywords). As will be discussed in Section 1.2.1, metadata are key to an efficient and effective access to multimedia content.

The term social media refers to the multimedia content contextualized in social networking (e.g., Facebook and Twitter) and content sharing (e.g., Flickr, YouTube and blip.tv) websites. While curated and professionally generated multimedia content collections increasingly also search the proximity of social media platforms, a large portion of the multimedia content found on such platforms is user-generated. Approximately 4.5 million images, captured around the globe are uploaded to Flickr daily [3] and in many parts of the world the density of captured images is already big enough to enable new categories of applications solely relying on user-generated content. For example, we conjecture that the images contributed by the Flickr user community could be used to complement information available in the official, publicly available cultural heritage portals, by e.g., providing a tourist with an overview of a geographic area surrounding a particular monument. To illustrate the potential for such tourist applications, Fig. 1.2 shows in logarithmic scale the distribution of geo-referenced



Miyajima: Floating torii gate

This beautiful floating torii gate is located in front of the Itsukushima shrine on Miyajima island. The island can be reached by ferry from Hiroshima.

Comments and faves

By mr. no-name
No real name given + Add Contact

This photo was taken on May 30, 2009 in Hiroshima Prefecture, JP, using a Canon PowerShot SX10 IS.

This photo belongs to

mr. no-name's photostream (1)

Tags

japan • miyajima • island • itsukushima • shrine • torii • gate

License

Some rights reserved

Privacy

🟩 This photo is visible to everyone

Figure 1.3: An example Flickr photo associated with e.g., user-generated title, description, keywords, comments as well as the automatically captured metadata and information on author.

Flickr images captured within a radius of 1km from each monument in The Netherlands and available under a creative commons license. The information used to produce the map was collected via Flickr API in January, 2012. The median number of images per location is 236 and it reaches the number of 24327 for a geographic area around a monument in Amsterdam.

In a social media environment, an image or a video is commonly associated with rich metadata, ranging from those automatically added by a content capturing device (or a social media website) and user generated textual annotations (title, tags, comments), to information about users and their social network. An example Flickr image and the associated metadata are shown in Fig. 1.3. While, previously, the users were generally not inclined to annotate their images and videos, the social dimension of the content sharing websites appears to be a strong tagging incentive [5]. Compared to e.g., the professional video collections, in which a video is annotated by at most a single professional archivist, in social media environments, the textual annotations are frequently generated by multiple users. In addition, comments posted by the users may explicitly or implicitly include information about the sentiment evoked by the content. Such richness of information may be beneficial as it may, for example, improve the annotation robustness in cases when e.g., multiple users assign the same tag to a multimedia item, but it may lead to the problems as well.

Namely, compared to those generated by the professional archivists, user generated annotations are often noisy, imprecise [40] and sometimes completely irrelevant to the content [76], which poses a significant challenge to the development of multimedia information retrieval solutions relying on such annotations.

1.2 Search and summarization

As indicated earlier in this chapter, the research reported in this thesis focused on two aspects of multimedia information retrieval: search and summarization. In this section we describe these terms in more detail and introduce some of the underlying terminology used further in the thesis.

1.2.1 Search

The goal of a search mechanism is to produce a list of results satisfying the user's information need specified by the query. Depending on the type of information stored in the collection and the search interface, the query may consist of text, images, music, video segments or a combination of those. For example, as illustrated in Fig. 1.4, to produce a list of images of the Arc de Triomphe in Paris using an image search engine, one may simply issue a query text "Arc de Triomphe, Paris", upload a photo of the arc or provide its geo-coordinates. At the moment, various commercial image search services, such as e.g., Google Images[2], support both text queries as well as the querying by an image example.

The essential element of the search mechanism is *relevance ranking*. Depending on the degree to which it satisfies a particular information need, an item in a data collection may be judged as relevant or irrelevant. Consequently, the quality of the search result is normally determined by the number of relevant items ranked high in the results list produced for a given query. One of the main challenges in designing the search mechanism is therefore defining a reliable method for measuring the relevance of the item to the query. This relevance is typically determined by computing the similarity between an item and the query. To be able to efficiently assess this similarity, both the query and the items in the collection should be represented in the same form, as illustrated in Fig. 1.4. The process of generating such representation is also known as indexing. At the indexing

[2] http://images.google.nl/

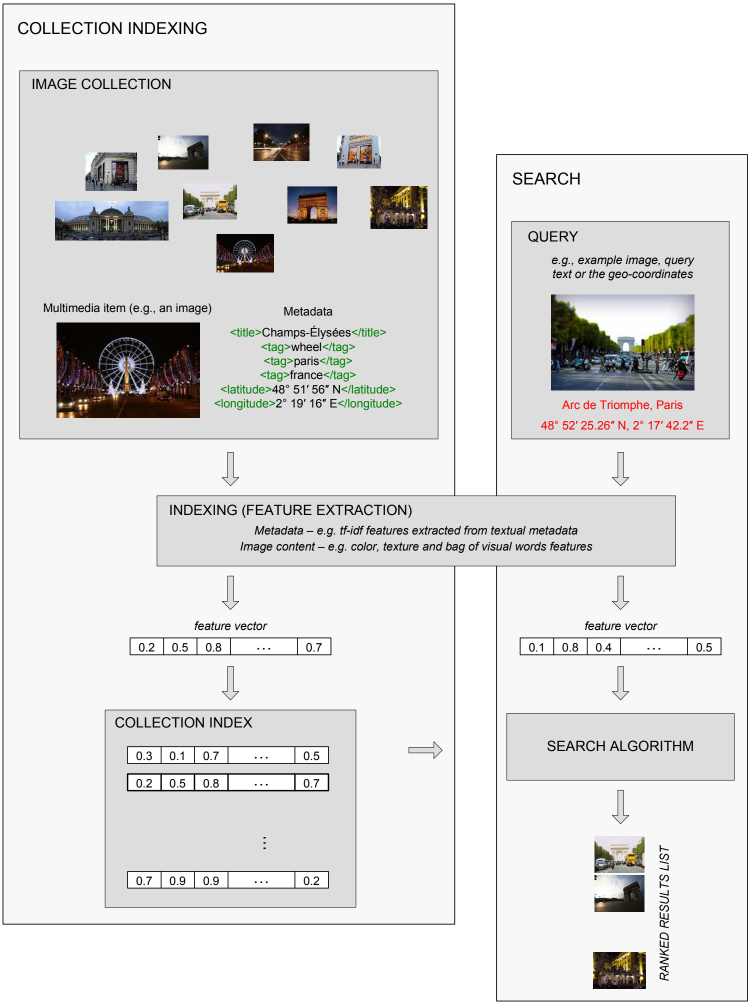

COLLECTION INDEXING
IMAGE COLLECTION
Multimedia item (e.g., an image)
Metadata
<title>Champs-Élysées</title>
<tag>wheel</tag>
<tag>paris</tag>
<tag>france</tag>
<latitude>48° 51′ 56″ N</latitude>
<longitude>2° 19′ 16″ E</longitude>
SEARCH
QUERY
e.g., example image, query text or the geo-coordinates
Arc de Triomphe, Paris
48° 52′ 25.26″ N, 2° 17′ 42.2″ E
INDEXING (FEATURE EXTRACTION)
Metadata – e.g. tf-idf features extracted from textual metadata
Image content – e.g. color, texture and bag of visual words features
feature vector
0.2 0.5 0.8 ··· 0.7
feature vector
0.1 0.8 0.4 ··· 0.5
COLLECTION INDEX
0.3 0.1 0.7 ··· 0.5
0.2 0.5 0.8 ··· 0.7
⋮
0.7 0.9 0.9 ··· 0.2
SEARCH ALGORITHM
RANKED RESULTS LIST

Figure 1.4: Illustration of indexing and search in an image collection.

time, a multimedia item is represented with a vector (or a set of vectors) of numerical values, commonly referred to as the *feature vector*.

Over the years, a plethora of features for content representation have been proposed. For example, a text document may be represented based on the frequency of terms occurring in it [52], while e.g. color or texture features may serve to represent an image [47]. However, as will be discussed in more detail in Section 1.3, the features that can be automatically extracted by a machine (e.g., color or texture features in case of images) as well as the content understanding based on them are rarely matching the level of a complex human interpretation. Therefore, various metadata are automatically or manually generated to facilitate the improved content indexing and retrieval. In practice, those metadata, depending on their type and properties, can be used directly without the additional processing or they can be indexed in a similar manner as the content of the same modality. For example, geo-coordinates automatically associated with the content by a capturing device, can be used directly to focus the search on a particular geographic area only, while the textual metadata inserted by a human such as e.g., title, tags and description can be indexed as a text document.

Numerous methods have been proposed to automatically generate textual metadata for an image or a video. These methods typically fall under the category of *visual concept detection* [35, 96] and aim at identifying objects, events or settings depicted in an image or the visual channel of a video. Typical examples of visual concepts are *person*, *crowd*, *vehicle*, *explosion*, *indoor* and *outdoor*. An image or a video may, however, also be represented using the features of multiple modalities, extracted from both the content and the metadata, which may be combined together to obtain better search results.

While the quality of the search result directly depends on the quality and effectiveness of feature representation, in general it is also highly dependent on the sophistication of the ranking method. Besides the straightforward ordering of the results according to their computed relevance values, various reranking algorithms [32, 79, 104] have been proposed that aim at refining the initial results list using the additional information available in the search process.

The quality of the results list is judged with an appropriate evaluation metrics, some of the commonly used being precision (ratio between relevant and the total number of items in the results list), recall (ratio between relevant items in the results list and the total number of relevant items in the collection) and average precision (a measure combining the precision and recall) [52]. In situations when several retrieval algorithms are available, the ability to automatically select the one producing the highest quality results list for a given query may improve the overall retrieval performance significantly. The process of predicting the quality of a results list for a given query is known as query performance prediction [15, 88, 116].

1.2.2 Summarization

Interacting with the content of a multimedia collection does not necessarily need to be done through a search interface. Browsing, for example, makes the information retrieval possible in cases when e.g., the users do not have any particular information need or are unable or unwilling to formulate it as a query. In a typical browsing scenario, a user is presented with a limited number of items selected according to a predefined criterion and the collection is explored in an interactive and "curiosity-driven" fashion. For instance, a user who missed a soccer match may be interested in viewing a short summary, in which only a limited number of video fragments are selected carefully to provide optimal insight into the course of the match. To enable such interaction with the collection, the effective methods for multimedia summarization are needed. To create a soccer match summary, a video may be represented by e.g., a curve showing the changes in users' excitement, from which the most exciting segments can be sampled for inclusion in the summary [23, 95]. In the other examples, one or more text documents may be summarized with a shorter text conveying the same message [66] and a collection of images may be summarized with a smaller image set [81]. In general, prior to summarization, a multimedia collection should be indexed, which may be performed as described in the previous section. Depending on the content type stored in the collection and a particular summarization purpose, a summary may consist of various content forms such as e.g., text, images and video segments. Although, as will be discussed in detail in chapters 3 and 4, the evaluation of summaries generally appears to be a significantly more complex problem than the evaluation of search results lists, they are typically based on the same general principles. Namely, first the relevant items are identified according to the predefined criteria and then the quality of a summary is evaluated based on the number of relevant items included.

1.3 Thesis scope and layout

The essential parameter of both search and summarization is the relevance criterion deployed to determine the ranking or steer the filtering of the collection items, respectively. As illustrated by the examples in Fig. 1.5, it can be drawn from a broad range of criteria defined at different semantic levels. Namely, the relevance of images or videos from a collection may be estimated e.g., (a) based on their basic visual composition (e.g., "find

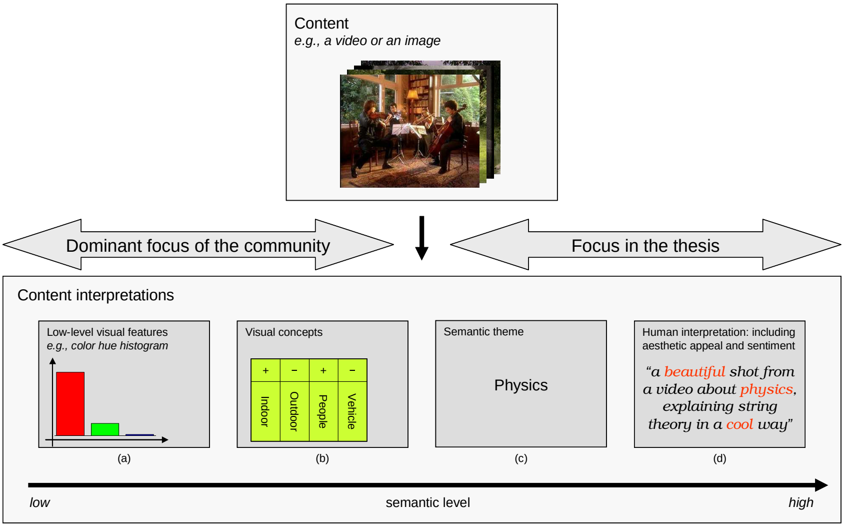

Content
e.g., a video or an image
Dominant focus of the community
Focus in the thesis
Content interpretations
Low-level visual features
e.g., color hue histogram
Visual concepts
+
−
+
−
Indoor
Outdoor
People
Vehicle
Semantic theme
Physics
Human interpretation: including aesthetic appeal and sentiment
"a beautiful shot from a video about physics, explaining string theory in a cool way"
(a)
(b)
(c)
(d)
low
semantic level
high

Figure 1.5: Content interpretations at different semantic levels: (a) low-level visual features, such as e.g., color histograms, (b) visual concept detectors, (c) semantic theme and (d) human interpretation, which includes high level attributes such as aesthetic appeal and sentiment.

me all images or video segments having a similar color distribution as the query image"); (b) based on the visual concepts detected in an image or the visual channel of a video (e.g., "find me the images or video segments of people playing string instruments in the indoor setting"); (c) based on what a video is actually about (e.g., "find me the videos covering the same topic - physics") or, eventually, (d) based on the user's information need defined at a much higher semantic level and comprising various criteria such as e.g., "*aboutness*", *aesthetic appeal* and *sentiment*.

The attention of the MIR research community has focused mainly on the relevance criteria defined at a lower semantic level (e.g., examples (a) and (b) in Fig. 1.5), as their addressing still poses significant challenges. However, little has been done on exploring the semantically more complex criteria characterizing the second two examples. The main research question underlying the work reported in the thesis can therefore be formulated as follows:

Can video search and visual summarization be performed based on the relevance criteria defined at a higher semantic level?

To answer the question, we approach it from different perspectives, analyzing different types of multimedia collections, information access paradigms and use-cases. As discussed in previous sections, properties of multimedia collections may vary significantly depending on whether e.g., content was produced and annotated by the professional archivist or the users of a content sharing platform. Therefore, we focus here on practical use-cases associated with two substantially different environments, namely a large professional video archive (Chapter 2) and a social media website (chapters 3 and 4). Further, we concentrate our study on the "aboutness" as the relevance criterion and while in Chapter 2 we address the question from the search perspective, in chapters 3 and 4 our focus is on summarization. Additionally, in Chapter 4 we also focus on the subjective summarization and summary evaluation criteria, where aesthetic appeal and sentiment have been found to play an important role.

1.3.1 Video search in a professional collection setting

In **Chapter 2**, we adopt a use-case of an archivist at a professional video archive searching for the videos about a given topic in order to re-use, modify or otherwise exploit them. We refer to the topic, or the general subject matter of the video, as the *semantic theme*. Examples of the queries defined at the level of the semantic theme in such search scenario may include *economy*, *politics*, *paintings* and *scientific research* among many others. We assume the most challenging and the most realistic scenario, in which videos are not labeled in any way and in which the search is based on the information that can be extracted from the multimodal video data only. The main research question underlying the work reported in Chapter 2 can therefore be defined as follows:

How to facilitate video search at the level of semantic theme by relying on the visual and spoken content of the video only?

The biggest obstacle we face in Chapter 2 when seeking to provide an answer to this question, is a relatively weak relation between the visual content of a video and its semantic theme. As illustrated by Fig. 1.5, the visual content of a video is only loosely related to its semantic theme - physics. Furthermore, the visual features that can be extracted from the video, such as e.g., color, texture or even visual concepts, are insufficiently indicative of its semantic theme. Additionally, as the semantic theme is an attribute of an entire video or a video segment of significant length, our retrieval unit is considerably longer than in the case of most related works in the field, which typically address queries specified at a lower semantic level with a stronger reference to the visual channel (e.g., queries referring to the objects or personages appearing in the videos) and targeting individual shots or scenes. The essence of the proposed method then lies in the idea to aggregate the outputs of visual concept detectors operating at the lower semantic level and applied to the video shots, to create a representation of an entire video, capable of encoding the information about its semantic theme. Finally, to effectively deal with a high diversity of semantic themes, the proposed retrieval framework incorporates the query performance prediction, making possible selection of the most appropriate retrieval algorithm for a given topical query. While the intermediate results of our research on the topic were published in [79, 83, 84, 85, 86] the final results were reported in [88] that we adopted as Chapter 2 of the thesis.

1.3.2 Visual summarization in a social media setting

In **Chapter 3** we move from the setting of an unlabeled, professional multimedia collection to the information-rich social media. We address the use case where a visual summary of a given geographic area needs to be generated automatically from the available user-contributed images. The visual summary is expected to illustrate not only the most dominant landmarks, but instead all relevant aspects of the geographic area, such as e.g., landmarks, museums, stores and restaurants. Since the imposed relevance criterion targets a deeper meaning of the images, it can be considered equivalent to the aboutness criterion from Chapter 2. However, we conjecture that the establishment of semantic relations between images to implement such relevance criterion can be improved significantly if we go beyond analysis of their visual content and make use of heterogeneous information associated with them in a social media setting, such as e.g., user generated metadata (title, tags, comments) and the information about users' interactions with the images and their social network. The research question underlying the work reported in Chapter 3 can therefore be formulated as follows:

How to maximize the quality of a visual summary, given the available social media information resources?

While, again, the essence of the proposed method was first published as [80], the detailed method description appeared in [81], which we adopt as Chapter 3 of the thesis.

Evaluating the quality of the obtained summary is not a trivial task as the absolute reference (ground truth) is hard, if not impossible, to acquire. The method reported in Chapter 3 was therefore evaluated simply in terms of the “necessary condition” required to be fulfilled in the envisioned summarization use case, namely to yield a good geographic coverage of a selected area.

This, however, immediately triggered our research into automatically generating the visual summaries satisfying the “sufficient condition” for being judged as high quality by the users, by, namely, automatically selecting the images the users would select if the summaries were generated manually. Only in this case, the summary can be said to maximally reflect the interests and needs of the user. This motivated the research question addressed in **Chapter 4**:

Is it possible to automatically identify the images that a user would consider suitable for creating a visual summary?

The crowdsourcing platforms recently emerged as the time and cost efficient tools for completion of tasks requiring human intelligence. We first investigate their potential for getting the insight into how humans perform visual summarization. We demonstrate that modeling image selection criteria and their interplay requires an unorthodox and heterogeneous set of image features, based on the analysis of their content, context, popularity in a social network, aesthetic appeal as well as the sentiment they evoke. The outcome of our proposed image selection approach is a list of images sorted according to the likelihood that they would be selected for the visual summary by a human. Finally, we investigate the possibilities of automatically evaluating the quality of image sets based on the human-created references, a problem which received insufficient attention in the research community. The work presented in Chapter 4 was reported as [82].

Chapter 2

Semantic-theme-based video retrieval

In this chapter we propose a novel approach to video retrieval at the level of semantic theme. The approach is based on the query performance prediction principle, which we deploy to choose the best retrieval results given a topical query, video collection and the available resources. We demonstrate that by jointly utilizing the automatic speech recognition and visual concept detection, video retrieval at the level of semantic theme can be efficiently performed even in a challenging environment of an unlabeled video collection.

This chapter is published as: Stevan Rudinac, Martha Larson, Alan Hanjalic. Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval. *International Journal of Multimedia Information Retrieval*, vol. 1, pp. 263-280, 2012.

2.1 Introduction

In this chapter we address the problem of video retrieval at the *semantic theme* level, where semantic theme refers to a general subject matter (topic) of a video. The query is formulated to encode a topical information need of the user and the retrieval system is expected to return videos that treat relevant subjects. Examples of such "topical" queries are *court hearings*, *youth programs*, *archaeology*, *celebrations*, *scientific research*, *economics*, *politics* and *zoos*.

Semantic themes come in a variety of abstraction levels and degrees to which they are visually constraining. In practice, a set of semantic themes might include video genres in a more traditional sense [7, 108] or the semantic labels assigned by archivists in professional libraries. They can, however, also correspond to the categories used in online content sharing portals, such as YouTube[3] and blip.tv[4].

A high level of inter-annotator agreement observed in professional digital libraries indicates that humans easily agree on the semantic theme of a video. Although it is not obvious where this inter-annotator agreement comes from, we hypothesize that both the visual and spoken content channel (ASR output) provide valuable information in this respect. While support for this hypothesis in the case of the spoken content channel was provided in our previous work [85], our goal in this chapter is to investigate the potential of the visual channel to help retrieve videos using topical queries.

On a first sight, the information in the visual channel may seem rather unreliable as an indicator of the general topic of a video. As shown in the examples in Fig. 2.1, frames extracted from different shots of a video covering the topic *youth programs* are characterized by highly diverse visual content that also does not directly connect a shot to the topic specified by the query. However, in view of the fact that the visual channel is used to complement or illustrate the topic of a video, it should not be surprising if the same key elements of the visual content, such as objects or parts of the scenery, appear in a large number of video clips covering the same semantic theme. Observed from this perspective, the visual content across different video shots in Fig. 2.1 may indeed be found consistent at a particular level of content representation, namely at the level of visual concepts. Here, our

[3] www.youtube.com

[4] blip.tv

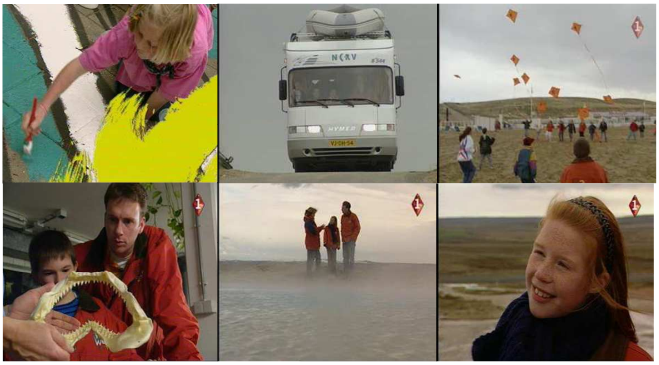


Figure 2.1: Keyframes of shots from a video in the TRECVID 2009 collection that is relevant to the semantic theme *youth programs*. The visual content of the shots contains information only weakly related to what the entire video is actually about.

definition of a visual concept corresponds to the definition adopted in the TRECVID benchmark [63] and represented by the ontologies such as e.g., the LSCOM [56]. Typical examples of visual concepts are *vehicle*, *meeting*, *outdoor*, *waterscape*, *flag* and - as in the case of the examples in Fig. 2.1 - *people*. In the same way, videos about *court hearings* could be expected to include many indoor scenes in courtrooms, while videos about *zoos* could be expected to depict animals significantly more often than other visual concepts. Videos about *celebrations* and *politics* typically contain shots involving people, but with different occurrence patterns: frequent appearance of larger groups of people might be more typical in case of celebration, whereas a video about politics would include more shots of individual people (e.g., taken during interviews with individual politicians).

In view of the above, the information on visual concepts should not go unexploited for the purpose of retrieving videos based on semantic themes. While this information remains insufficient to link a video directly to a topical query, we foresee a large value of this information in its ability to help determine whether two videos are similar in terms of their semantic themes. As we also conjecture that the visual concept detectors have a potential to encode information about stylistic features related to e.g., television production rules [55], their value for determining video similarity may expand across a broad range of semantic themes defined at various abstraction levels.

We propose in this chapter a retrieval approach that consists of the following two steps:

- Building a video representation that is suitable for assessing similarity between two videos in terms of their semantic themes and that is based on aggregating the outputs of visual concept detectors across different shots of a video, and

- Query expansion selection (QES) that responds to topical queries and that is based on the query performance prediction (QPP) principle (e.g., [15, 116]). Here, the proposed video representation serves as input into query performance indicators, which evaluate various results lists produced by different query modifications.

The list with the highest estimated performance is then adopted as the best possible search result given a topical query, video collection, available search mechanisms and the resources for query modification.

The main research questions we address in this chapter are

- To which extent can the proposed QES retrieval approach outperform a baseline system that solely relies on the spoken content channel?

- For which categories or abstraction levels of semantic themes does the QES approach work well and what reasons of failure can be inferred for semantic themes for which the approach fails?

- Is it possible to obtain a more reliable prediction through combining concept-based indicators and text-based indicators of query performance?

We first explain the rationale and outline the contribution of our retrieval approach in Section 2.2, while in Section 2.3 we provide an insight into the state-of-the-art in the main technologies underlying this approach. Then, we introduce the two main approach steps listed above, namely building the video representation that we refer to as *Concept Vector* (Section 2.4) and designing the QES retrieval framework utilizing this video representation (Section 2.5). Sections 2.6, 2.7 and 2.8 are dedicated to the experimental evaluation of our approach. Sections 2.6 and 2.7 address the first two research questions mentioned above, while the third research question is addressed in Section 2.8. The discussion in Section 2.9 concludes the chapter.

2.2 Approach rationale and contribution

We base our approach on the same rationale that is underlying general QPP approaches [15, 30, 116] and which builds on the clustering theorem [107] stating that closely related documents tend to be relevant to the same request. In our approach, for analyzing the relatedness between videos in terms of a semantic theme, we rely on the discussion in Section 2.1 and propose a video representation that exploits general distribution patterns of a large set of visual concepts detected in a video. Hereby, we do not assume that a special set of visual concepts must be detected for a given video collection. In other words, our approach does not require the assurance that the concept set used provides complete semantic coverage of the visual content of the collection. The possibility to work with a general set of visual concept detectors makes our retrieval approach unsupervised and therefore opens a broader search range than in the case of supervised alternatives. Examples of such alternatives are the approaches that learn or otherwise generate mappings between specific visual concepts and semantic themes. Such approaches, which have been studied for shot-level retrieval, cf. [34, 97], face the challenge of collecting a sufficiently large and representative set of visual concepts, particularly daunting for never-before-seen topical queries and being rather sensitive to the quality of visual concept detectors. Furthermore, as discussed in more detail in Section 2.3.3, such approaches are commonly tailored for TRECVID-like queries, differing from semantic themes in their strong reference to the visual channel of the video. In addition, since statistical information is collected over a large set of concept detectors, our approach is less sensitive to noise in the individual detectors.

Different results lists serving as input to query performance prediction are obtained for different query expansions created by adding additional terms to the original query. Query expansion (see Section 2.3.1 for more information) is widely deployed in the field of information retrieval (IR) in order to enrich the original query so as to provide a better match with documents in the target collection. In particular, it is known to increase recall [52]. In the area of spoken content retrieval, query expansion is often used [38, 113] where it also compensates for errors in the speech recognition transcripts. The danger of query expansion is, however, that it may introduce inappropriate terms into the query, causing topical drift. Given an initial query text, a speech transcript of a video collection and a set of search results lists obtained for different query expansion methods and applied to the speech transcript, our QES approach controls the drift and selects the most appropriate query expansion method.

In our previous work [85], coherence indicators of query performance, exploiting pair-wise video similarities in terms of their spoken content, demonstrated the ability to improve retrieval at the semantic theme level within the proposed QES framework. In this chapter, we revisit and adjust this framework to first investigate to which extent a modification of these *text-based* coherence indicators into the indicators exploiting *concept-based* similarities between videos can lead to an improvement of the semantic-theme-based video retrieval within the QES framework. Then, we also investigate whether additional improvement could be achieved by combining text-based and concept-based indicators.

In addition to being the first work to address in depth the problem of semantic-theme-based video retrieval, the main novel technical contribution of our approach is an integration of the output of visual-concept detectors aggregated across the entire video and the output of automatic speech recognition, both known to be noisy. We will show that through such integration an overall improvement in retrieving videos using topical queries can be achieved, compared to several baseline approaches commonly used in the IR field. More specifically, we will demonstrate that for a given query our concept-based query performance indicators are indeed effective in selecting the best out of available search results lists. Finally, we will show that a simple combination of concept-based indicators with the text-based alternatives might significantly improve performance in terms of mean average precision (MAP) and that, more importantly, a combined coherence indicator selects the optimal results list in over 35% queries more than state-of-the-art text-based indicators.

2.3 Related work

2.3.1 Query expansion

A common problem in information retrieval is a mismatch between vocabularies of the query and the collection being queried. This problem is often addressed by expanding the query using, for instance, pseudo-relevance feedback or thesauri. Query expansion can be particularly helpful in the case of spoken content retrieval in which speech recognizer errors and particularly errors caused by words spoken in the recognizer input, but missing in the recognizer vocabulary frequently occur. It is sometimes difficult to separate the improvement contributed by the expansion itself from the error compensating effects, but overall query expansion is known to yield improvement [38, 113]. For example, recognizer error occurring for the original query term *excavation* might be compensated by expanding the query with additional related terms, such as *digging*, *archaeology*, *archaeologist* and *artifacts*, which are potentially correctly recognized. Although proper query expansion may generally improve the retrieval results, it also introduces the danger of a topical drift [45], the tendency of expanded query to move away from the topic expressed by the original query.

2.3.2 Query performance prediction

Topical drift can be controlled by appropriate query performance prediction applied to decide whether a query should be expanded and how [27]. In particular, our work is related to methods for post-retrieval query prediction, i.e., methods that use results lists returned by an initial retrieval run as the basis for their performance prediction. In [15], query prediction uses the Kullback-Leibler divergence between the query model and the background collection model (clarity score). Yom-Tov et al. [116] proposed efficient and robust methods for query performance prediction based on measuring the overlap between the results returned for a full query and its sub-queries.

Recently, a coherence-based approach to query prediction has been proposed [30]. This approach measures the topical coherence of top documents in the results list in order to predict query performance. The approach is low in computational complexity and requires no labeled training data. Further, the coherence-based approach is appealing because it goes beyond measuring the similarity of the top documents in a results list to measuring their topical clustering structure [31]. The coherence score is thus able to identify a results list as high-quality even in the face of relatively large diversity among the topical clusters in the top of results list.

In our recent work [85], we demonstrated the performance of the coherence score defined in [31] and two light-weight alternatives for the task of text-based QES. Subsequently, we carried out initial work, reported briefly in [84, 86], which established the potential of coherence score to be useful for multimodal QES. In this chapter, we present the fully developed version of that initial approach including automatic generation of Concept Vectors for video representation, combining the proposed text-based and concept-based query performance indicators and validation on a large dataset.

In [103], an approach to performance comparison of web image search results has been proposed. The underlying ideas, including assumptions on density of relevant and non-relevant images and their pairwise similarities place this approach into the group of coherence-based approaches. However, it requires training and relies on preference learning, which could eventually reduce applicability to unseen queries. In addition, the set of queries used in the experiments indicates a strong reference to the visual channel and it remains unclear whether the approach could be applied for multimedia information retrieval at a higher semantic level, especially since the models were built based on low-level visual features only.

2.3.3 Multimodal video retrieval

Since a video conventionally consists of both a visual and audio track, multimodal approaches are clearly necessary in order to exploit all available information to benefit video retrieval. Our QES approach bears closest resemblance to reranking approaches, which use visual information to refine the initial results returned by spoken content retrieval [32, 79, 104]. However, there are important differences between QES and reranking. First, reranking approaches are restricted to reordering the initial results list - there is no mechanism that allows visual features to help admit additional results. Second, reranking methods are static and therefore known to benefit some queries and not others [32, 79, 104], while our QES approach adapts itself to queries. It attempts to maximally exploit the available information to select the best results list per query.

Another important difference between the work presented here and the previous work is the type of the retrieval task. As noted in the introduction, semantic-theme-based video retrieval involves retrieving video according to its subject matter. Typical semantic theme (topical) queries are thus defined at a higher abstraction level and therefore substantially different from conventional TRECVID queries, which include named persons, named objects, general objects, scenes and sports (cf. [28]). TRECVID-type queries are strongly related to the visual channel and may not be actually representative of the overall topic of the video. This difference is reflected in the size of the retrieval unit. Unlike the majority of approaches that address video retrieval at the shot level (e.g., [32, 59, 98, 104]), we consider entire videos as retrieval units. Our decision to move beyond shot-level retrieval is guided by the reasoning that a semantic theme is an attribute of either an entire video or a video segment of a significant length. We also believe that in many real-world search scenarios, e.g., popular content sharing websites, such as *YouTube* and *blip.tv*, users are actually looking for the entire videos to watch and that clips or segments must be of a certain minimum length in order to satisfy users' information need. While there has been little effort in the past that targeted video retrieval beyond the level of individual shots, recently, a story-level video retrieval approach was proposed that retrieves news items containing visually relevant shots [4]. Although relevance is not assessed with respect to the semantic theme, we mention this approach here because it is similar to our own regarding a relatively large retrieval unit and also the use of language models built over the concept detector output.

The increasing awareness of the need to address queries at a higher abstraction level than e.g., LSCOM, can also be observed from the reformulation of a TRECVID search task, which was renamed to *known item search task* in TRECVID 2010 [63] and which included a limited number of theme-based queries, as well as a new video-level retrieval evaluation metric.

2.4 Building concept vectors

In this section, we present our approach for automatically creating Concept Vectors, visual concept-based representations of videos that are used to calculate similarities between videos that capture resemblances in terms of a semantic theme.

2.4.1 Making use of incomplete sets of noisy visual concept detectors

Since the relation between the semantic theme of a video and its visual content is potentially weak, the problem of successfully utilizing the visual channel for the purpose of query performance prediction appears to be rather challenging. In view of the discussion in Section 2.1, we believe that the intermediate representation at the level of visual concepts could lead to a solution for this task. Like words in the speech channel, concepts in the visual channel can be said to reflect the elements of the thematic content of the video.

A critical challenge to making effective use of visual concepts is the relatively low performance of state-of-the-art visual concept detectors. As an example, the performance in terms of mean average precision (MAP) of the best performer in “Concept Detection” and “Interactive Search” tasks of TRECVID 2009 was below 0.25 [98]. Our approach is based on the insight that in spite of a relatively poor performance and noisiness of individual visual concept detectors at the shot level, aggregating the results of concept detections across a series of shots could help reduce the influence of this noise and still provide the basis for a reasonable video-level representation in the use context addressed in this chapter.

The question has been raised in literature of how many and which concept detectors would be required to sufficiently cover the entire semantic space for the purpose of effective video retrieval in a general use case [29]. Although, ideally, as many concept detectors as possible should be available in order to be able to handle enormous diversity of visual content and address a broad range of video search requests, the reality is that the set of available concept detectors will always be limited and not necessarily representative for every content domain. We hypothesize, however, that availability of the optimal visual concept set for a given use case is not critical for successful deployment of our approach, provided that mechanisms are developed to determine which particular concepts from the available concept set are more informative to be applied on a particular video collection.

Based on the above two hypotheses, we approach automatic generation of Concept Vectors by starting from an arbitrary set of available visual concept detectors, analyzing their output and selecting the most representative (informative and discriminative) visual concepts. Technical steps of this approach are described in more detailed in the subsequent parts of this section.

2.4.2 Concept-based video representation

To create our Concept Vectors, we follow the general process illustrated in Fig. 2.2 in which we draw an analogy to the conventional information retrieval and consider visual concepts as terms in a video “document”. In this process, we aim at representing a video v from a collection V using a vector $\mathbf{x}_v$ defined as

$$\mathbf{x}_v = [x_{1v}, x_{2v}, \ldots, x_{|C|v}]^\top \tag{2.1}$$

where x_{cv} is the weight of the concept c in video v, C is a general set of visual concepts and $^\top$ is the transpose operator. The weight x_{cv} serves to indicate the importance of the concept c in representing the video v. In the

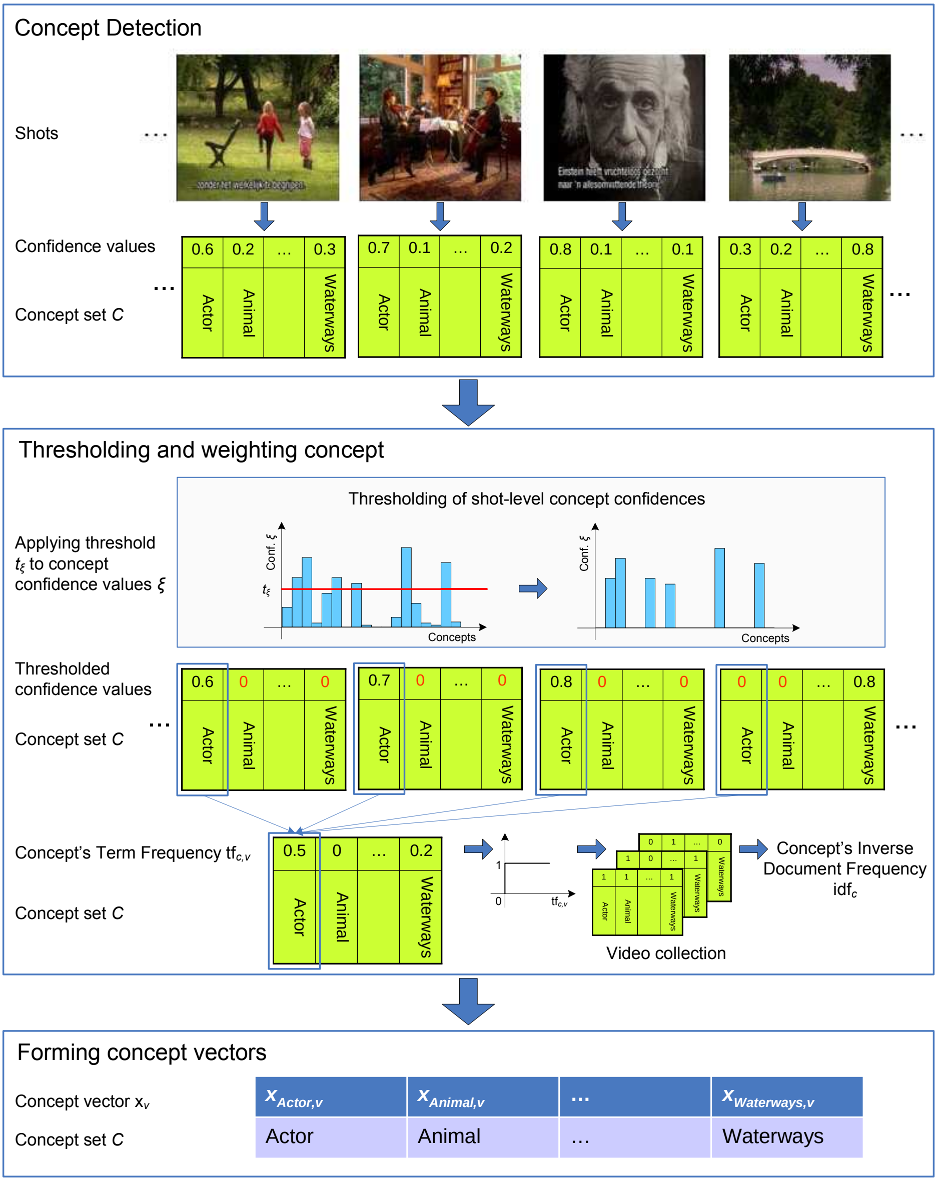


Figure 2.2: Illustration of our approach to concept-based video representation starting from a general concept set C. Final concept vectors $\tilde{\mathbf{x}}_v$ are created based on the subset $\widetilde{C}$ of selected concepts, as explained in Section 2.4.3.

conventional information retrieval, this importance is generally expressed as a function of the $\text{tf}_{c,v}$ (term frequency) and idf_c (inverse document frequency) [89], which reflect the number of occurrences of a term in a video and its discriminative power within the collection, respectively. The index "c, v" indicates that the TF component of the weight is specific for a video, while the index "c" reflects that the IDF component is computed over the entire collection.

When computing the $\text{tf}_{c,v}$ component of the weight, we take into account the fact that state-of-the-art concept detection systems [35, 96] usually output shot-level lists of confidence scores for the given visual concepts, rather than binary judgments. For this reason, we model the term frequency here by the sum of a concept's confidence scores taken from each shot of a video. In order to avoid bias towards videos containing more shots, we normalize the sum of confidence scores with the number of shots. Furthermore, recent works (i.e., [35, 96]) revealed that the values of visual concept confidence vary widely within the interval of $[0, 1]$, with low confidence values commonly indicating erroneous detection. Low confidence values effectively introduce a large amount of noise into the system, which will negatively bias the computation of idf_c. Therefore, we analyze the outputs of individual concept detectors and consider the reliable outputs only. In other words, we perform thresholding at the shot level to retain only those concepts in the representation that have substantial confidence scores. In our approach thresholding is an essential step also because, as revealed by our exploratory experiments, reliable indicator of term (concept) frequency is critical for reliably selecting a subset of representative concepts.

Taking into account the above considerations, we compute $\text{tf}_{c,v}$ according to the following expression:

$$\text{tf}_{c,v} = \frac{\sum_{j=1}^{N_v} \{\xi_{c,v,j} : \xi_{c,v,j} > t_\xi\}}{N_v} \tag{2.2}$$

Here, $\text{tf}_{c,v}$ is the normalized frequency of a concept c in video v, N_v is the number of shots in a video and $\xi_{c,v,j}$ is the confidence of the presence of a particular concept c in the shot j of video v as provided by the concept detector. The value of the threshold t_ξ we introduce for the purpose of denoising the output of the concept detectors is not critical if selected above a certain value. In our experiments, threshold values larger than 0.3

yielded insignificant difference in the performance.

$$\mathrm{idf}_c = \log \frac{|V|}{|\{v : \mathrm{tf}_{c,v} > 0\}|} \tag{2.3}$$

While $\mathrm{tf}_{c,v}$ represents the intensity of concept occurrence in a single video, idf_c (c.f. (2.3)) serves to incorporate the general pattern of visual concept occurrence within the entire collection. idf_c is computed by first dividing the entire number of videos in the collection with the number of videos in which the given concept is present and then by taking a logarithm of the quotient. Different ways of mapping $\mathrm{tf}_{c,v}$ and idf_c onto x_{cv} will be investigated experimentally in Section 2.7.

2.4.3 Concept selection

The goal of concept selection is to choose a subset of concepts from the available set C that are able to capture semantic similarities between videos. Concept selection can be seen as a feature selection problem known from the pattern recognition and information retrieval domain. Through years, many methods have been proposed to select features [102, 114], many of which are supervised and require prior training. Our previous work revealed a high positive correlation between the frequency of concept occurrence across the collection and its effectiveness in discriminating between videos based on the semantic theme [86]. In order to have our approach completely data driven and unsupervised, we introduce a method for concept selection based on a simple heuristics that involves computing of the *frequency*, *variance* and *kurtosis* of visual concepts in the video collection. As will be explained in Section 2.6.2, here we set $x_{cv} = \mathrm{tf}_{c,v}$.

Frequency. We conjecture that concepts that occur in many videos within the collection will be more helpful in comparing videos than those concepts appearing in only few videos (Fig. 2.3a). Then, the relative difference in the importance weights of such concepts can provide a basis for calculating similarity between two videos. For each concept c we compute freq_c by aggregating the concept counts a_{cv} across videos in which that concept appears:

$$\mathit{freq}_c = \sum_{v=1}^{|V|} a_{cv} \quad \text{and} \quad a_{cv} = \begin{cases} 1, & x_{cv} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

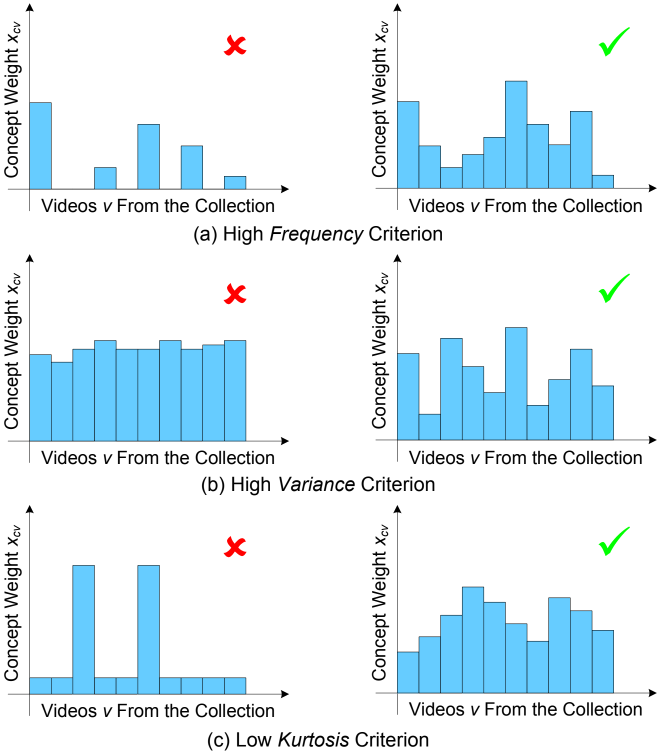


Figure 2.3: Illustration of *frequency*, *variance* and *kurtosis* criteria for concept selection. Distribution examples on the right show the desired behavior of frequency, variance and kurtosis for marking relevant visual concepts.

Variance. Selecting the frequent concepts only is not enough, since some frequent concepts might have importance weights distributed uniformly throughout the collection. In that case, the concept will not be discriminative for comparing videos. Therefore, we require these frequent concepts to also have a high *variance* (Fig. 2.3b) of their importance weights across the video collection as well:

$$var_c = \mathrm{var}(\mathbf{y}_c), \mathbf{y}_c = [x_{c1}, x_{c2}, \ldots, x_{c|V|}] \tag{2.5}$$

where $\mathbf{y}_c$ is the vector of weights of concept c in all videos in the collection.

Kurtosis. A high variance (2.5) might be the consequence of either infrequent extreme deviations or, preferably, frequent, but moderate variations of concept weights across the collection. To isolate the concepts with frequent but moderate variations, we focus on those concepts with a

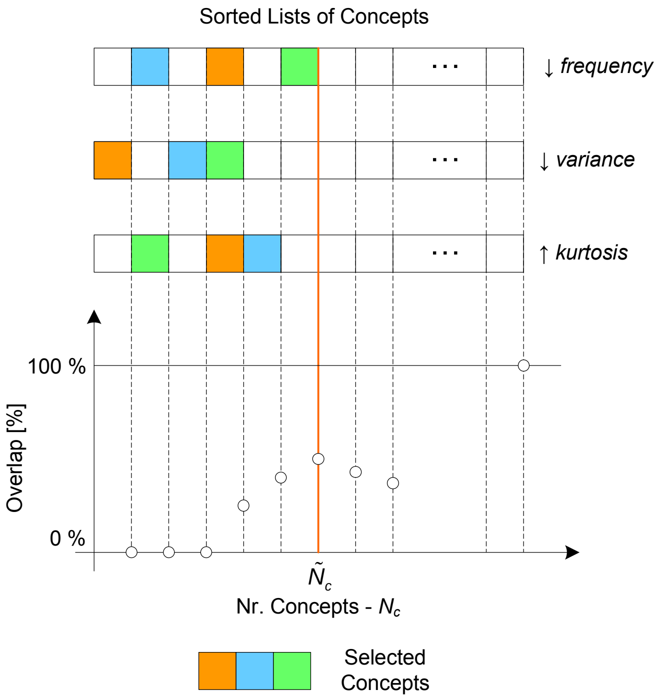


Figure 2.4: Illustration of the procedure for selecting concepts that satisfy the frequency, variance and kurtosis criteria.

low kurtosis. Kurtosis is a measure of "peakedness" of the probability distribution of a real-valued random variable (Fig. 2.3c). We compute $kurt_c$ of a concept using (2.6), where μ and σ are the mean and the standard deviation of the vector $\mathbf{y}_c$:

$$kurt_c = \frac{\sum_{v=1}^{|V|} (x_{cv} - \mu)^4}{(|V| - 1)\sigma^4} \tag{2.6}$$

As illustrated in Fig. 2.4, we produce three ranked lists by sorting the concepts according to the decreasing frequency and variance and increasing kurtosis in the collection. Then, we compute the percentage of the overlap between the three top-N_c lists for the increasing number N_c of top-ranked concepts. The process stops at $\widetilde{N}_c$, when the first dominant local maximum in the overlap value curve is reached (e.g., overlap of more than 70%), after which the concepts are selected that are common to all three top-$\widetilde{N}_c$ lists.

Prior to detecting local maxima, we smooth the overlap curve using a moving average filter, with the span parameter set to 10. The smoothing performed in this way helps reduce the influence of non-dominant local extrema and improves robustness of the concept selection approach. As will be shown in Section 2.6, the change in overlap with the increasing N_c remains largely consistent over different video collections and concept detection systems.

If we denote the three top-N_c lists of concepts sorted by *frequency*, *variance* and *kurtosis* as $\mathit{Freq}(N_c)$, $\mathit{Var}(N_c)$ and $\mathit{Kurt}(N_c)$, respectively, the selected set $\widetilde{C}$ of visual concepts can be defined as

$$\widetilde{C} = \mathit{Freq}(\widetilde{N}_c) \cap \mathit{Var}(\widetilde{N}_c) \cap \mathit{Kurt}(\widetilde{N}_c) \tag{2.7}$$

which leads to the “optimal” concept vector

$$\tilde{\mathbf{x}}_v = [\tilde{x}_{1v}, \tilde{x}_{2v}, \ldots, \tilde{x}_{|\widetilde{C}|v}]^\top \tag{2.8}$$

that serves as input for comparing videos in the subsequent query expansion selection step. In (2.8), $\tilde{x}_{cv}$ is the weight of concept $c \in \widetilde{C}$ in video $v \in V$ and $^\top$ is the transpose operator.

2.5 Query expansion selection

We approach the QES task from the data driven perspective, analyzing the collection being queried and the retrieval results list returned for the given query text. Fig. 2.5 illustrates our QES approach. The system makes an unsupervised online analysis of the results lists produced for the original query and multiple query expansions to decide whether the query should be expanded and if so, which of the alternative expansions would eventually yield the best results. An additional strength of our approach lies in the fact that we do not attempt to predict the retrieval performance (i.e., in terms of MAP) for each of the results lists (which usually requires prior training), but only compare the coherence of their top ranked results. We evaluate three coherence indicators for this purpose, which will be introduced in the remainder of this section.

2.5.1 Coherence indicator

The coherence indicator [31] is used to select the results list with the highest coherence among the top-N retrieved results. The indicator is computed

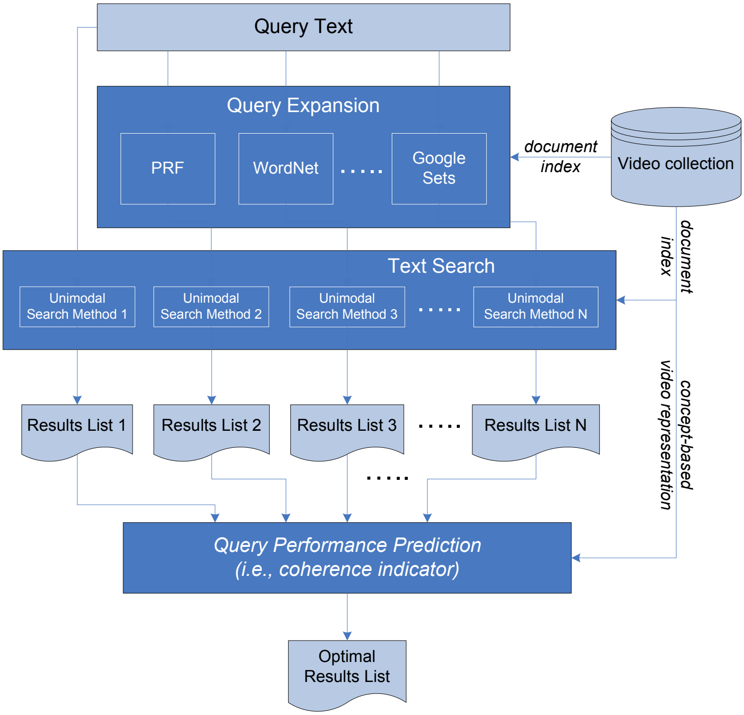


Figure 2.5: Illustration of our QES approach.

according to (2.9) as the ratio of video pairs in the top-N results whose similarity is larger than a threshold θ.

$$Co(TopN) = \frac{\sum_{u,v \in TopN, u \neq v} \delta(u,v)}{N(N-1)},$$
$$\delta(u,v) = \begin{cases} 1, & \mathrm{sim}(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v) > \theta \\ 0, & \text{otherwise} \end{cases} \tag{2.9}$$

The threshold θ is set as a similarity value between particularly close videos in the collection. The threshold choice will be further discussed in the experimental section. As a similarity measure $\mathrm{sim}(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v)$ we use the cosine similarity between the concept vectors (2.8) computed for videos u and v.

2.5.2 Max-AIS and mean-AIS indicators

The max-AIS and mean-AIS indicators [85] have been introduced as an alternative to the coherence score, because they do not need the reference to the video collection and make the decision based on the analysis of top-N ranked videos only. These indicators select the query expansion producing a results list in which top-N videos are characterized by high average item similarities (AIS) with their fellows. For video v AIS is computed according to (2.10).

$$AIS_v = \frac{\sum_{u \in TopN, u \neq v} \mathrm{sim}(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v)}{N-1} \tag{2.10}$$

Again, as a similarity measure $\mathrm{sim}(\tilde{\mathbf{x}}_u, \tilde{\mathbf{x}}_v)$ we use the cosine similarity between the concept vectors (2.8) computed for videos u and v. Max-AIS indicator takes the maximum AIS value of all top-N videos in the results list, while mean-AIS takes the average of the AIS values in the top-N.

2.6 Experimental setup

This section describes our experimental framework and gives the implementation details of our approach.

2.6.1 Datasets

The experiments are performed on two datasets, that are re-issues of the TRECVID 2007, 2008 and 2009 data made for the purposes of the "Tagging Task: Professional Version" offered for the MediaEval 2010[5] benchmark [44]. This benchmark also provided ground truth in the form of semantic theme labels assigned by professional archivists. The datasets are referred to as DS-1 and DS-2 and correspond to the MediaEval 2010 development and test dataset, respectively. Both datasets consist of the news magazine, science news, news reports, documentaries, educational programming and archival videos, provided by The Netherlands Institute for Sound and Vision (S&V)[6]. For the experiments we use both DS-1 and DS-2 to investigate generalization of our approach across datasets. Unless stated otherwise, we do not treat them as the development and the test set, but rather as two equal datasets.

[5] www.multimediaeval.org

[6] www.beeldengeluid.nl

Description of DS-1 dataset

The DS-1 dataset is a large subset (nearly all) of the TRECVID 2007 and 2008 datasets, which consist of 219 videos each (438 videos in total). In the process of creating the DS-1 dataset, the videos without a semantic theme label were removed. Further, the videos without the automatic speech recognition transcripts and/or machine translation were also discarded. This led to a dataset consisting of 405 videos. As the queries, 37 semantic theme labels assigned by the S&V archive staff were used. These labels were selected such that each of them has more than five videos associated with it. The list of labels was post-processed by a normalization process that included standardization of the form of the labels and elimination of labels encoding the names of personages or video sources (e.g., amateur video).

Description of DS-2 dataset

The DS-2 dataset is composed of videos from TRECVID 2009 dataset. Only videos (400 in total) that did not occur in TRECVID 2007 and 2008 were included. Again, the videos without a semantic label provided by the S&V have been removed. Further, the videos without the automatic speech recognition transcripts and/or machine translation were also discarded. This led to a dataset consisting of 378 videos. As the queries, a set of 41 semantic labels assigned by the S&V archive staff were used, defined as explained in the previous section. As with the DS-1 dataset, the list of labels was post-processed by a normalization process that included standardization of the form of the label.

As shown in tables 2.8 and 2.9, only 16 semantic labels are common to both DS-1 and DS-2 datasets, which serves to test the transferability of our approach to the never-before-seen queries. The performance stability across queries is analyzed in Section 2.7.6.

Query expansion methods

The query is modified using the following expansions:

- Conventional PRF (pseudo-relevance feedback), where 45 expansion terms are sampled from the automatic speech recognition transcripts of top-ranked videos in the initial results list produced for unexpanded query.

- WordNet expansion, by means of which the initial query terms are expanded with all of their synonyms. The average total number of terms in such expanded queries is 12 for DS-1 and 13 for DS-2.
- Google Sets expansion, in which the initial query is expanded with a certain number of items (words or multi-word phrases) that frequently co-occur with that query on the web. To control topical drift, we limit the number of expansion items to 15.

2.6.2 Visual concept detectors

Concept detector choice

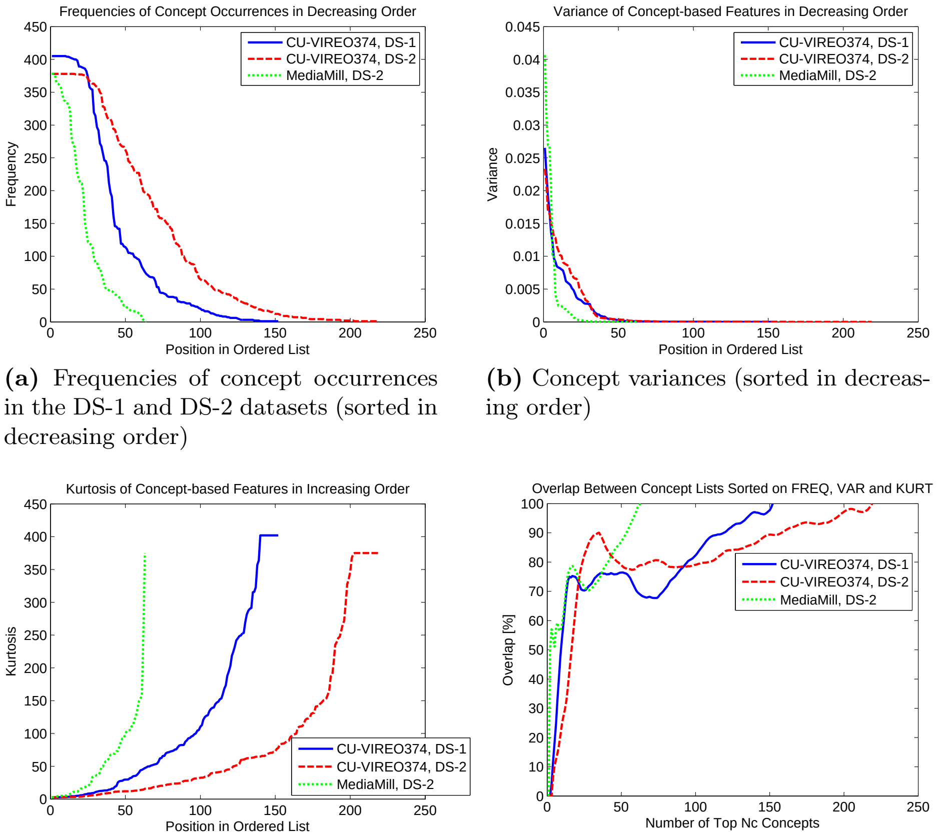
Videos from the DS-1 dataset are represented using CU-VIREO374 concept detection scores [35]. The system consists of 374 visual concepts selected from the LSCOM ontology [56]. To represent the DS-2 dataset we used (separately) both CU-VIREO374 and MediaMill [96] visual concept detection scores for the purpose of comparative analysis. MediaMill system consists of 64 concept detectors and at the moment when the experiments described here were performed, their outputs were publicly available for DS-2 dataset only.

Concept selection procedure

We now experimentally verify the feasibility of the methodology for selecting a subset of representative visual concepts for a given collection, which is based on the frequency, variance and kurtosis of the concepts, as described in Section 2.4.3. In the experiments reported in Section 2.7, TF weighting yielded a better performance than TF-IDF in the concept selection task and therefore for the computation of concept frequency, variance and kurtosis, c.f. (2.4), (2.5) and (2.6), we set here $x_{cv} = \mathrm{tf}_{c,v}$.

Figures 2.6a, 2.6b and 2.6c show the plots of frequency of concept occurrences, variance and kurtosis of $\mathrm{tf}_{c,v}$ throughout the video collection constructed using CU-VIREO374 concepts on the DS-1 dataset and CU-VIREO374 and MediaMill concept detectors on the DS-2 dataset.

The results shown in Fig. 2.6a indicate that some concepts are present in almost all of the videos in the collection with a significant confidence, while a large subset of concepts appear only in a limited number of videos. This observation holds for both the DS-1 and the DS-2 dataset and, surprisingly, both for CU-VIREO374 and MediaMill concept detectors (not



(a) Frequencies of concept occurrences in the DS-1 and DS-2 datasets (sorted in decreasing order)

(b) Concept variances (sorted in decreasing order)

(c) Concept kurtoses (sorted in increasing order)

(d) Percentage of overlap between the lists of concepts ordered according to frequency, variance and kurtosis

Figure 2.6: Illustration of concept frequency (a), variance (b) and kurtosis (c) across collections and the percentage of overlap between the lists of concepts sorted according to those criteria (d).

affected by the difference in number of concept detectors in both systems).

Also, as shown in Fig. 2.6b, a small subset of concepts has a high variance in the DS-1/DS-2 dataset, while a larger number of concepts show relatively uniform values across the collection. Similar observation can also be made for kurtosis (Fig. 2.6c). The goal of our concept selection procedure is to isolate a set of concepts that appear as high as possible in the concept ranking (i.e., as far to the left as possible in Fig. 2.6a-2.6c), meaning that they have high variance, high frequency and also low kurtosis. Finally, Fig. 2.6d shows the curves used to determine the length $\widetilde{N}_c$ of the ranked lists at which the set of selected concepts is generated as the overlap between the three lists. Supporting the illustration in Fig. 2.4, the curves indeed show clear local maxima at which $\widetilde{N}_c$ can be determined.

2.7 Experimental evaluation of QES

Through the experiments summarized in this section we seek answers to the following research questions:

- How does the proposed QES approach perform if videos are represented using the original concept vector (1), without refinement through the concept selection step (Section 2.7.1)?
- To which extent does the concept selection step improve the QES performance and under which conditions (Section 2.7.2)?
- Do the results generalize onto a new dataset and under which conditions (Section 2.7.3)?
- What is the impact of the quality of visual concept detection on the QES performance (Section 2.7.4)?
- Is the performance gain stable across queries (Section 2.7.6)?

We will address these questions first by working with the coherence indicator (Co) defined in (2.9). The performance of alternative indicators (mean-AIS and max-AIS) is then analyzed separately in Section 2.7.5. Furthermore, we will evaluate our approach in view of the above questions through a comparative analysis involving the best performing baseline approach. This reference approach is selected among the simple text-search baseline that uses speech recognition transcripts only and our three additional results lists produced using common query expansions (Section 2.6.1). The performance of these approaches in terms of MAP is shown in Table 2.1 for both DS-1 and DS-2 datasets. Note that the four results lists produced by these four baseline approaches are the ones that will be combined by our QES approach.

When interpreting the results, it is important to note that here we are not interested in improving the MAP in the absolute sense since MAP depends on the quality of the available baseline results lists. Instead, for

Table 2.1: MAP of the baseline and the query expansions used

Dataset	Baseline	PRF	WordNet	Google Sets
DS-1	0.2322	0.2619	0.1941	0.1271
DS-2	0.2381	0.2621	0.1867	0.1276

each query, we target to always choose the best results list, whatever MAP it has. In order to make a comparison with the theoretical optimum, we make use of "oracle" indicators, the hypothetical indicators that always choose the correct query expansion. Here we note that oracle indicators would achieve a MAP of 0.3082 and 0.3136 on the DS-1 and DS-2 dataset respectively. These numbers can be seen as the upper limits of the achievable performance of our QES approach. Throughout experiments we also compare the performance of our QES approach to the performance of the "Best Baseline", a baseline approach achieving the highest MAP for a given video collection.

Finally, the evaluation will take into account the influence of the main parameters of our approach:

- The threshold θ used for computing the Co indicator,
- top-N, the number of videos used for computing the Co, mean-AIS and max-AIS coherence indicators,
- the number of automatically selected visual concept detectors $|\widetilde{C}|$ and
- x_{cv}, that can be set to either TF or TF-IDF.

2.7.1 QES using all concepts

In the first query expansion selection (QES) experiment we use both CU-VIREO374 and MediaMill concept detectors. In Table 2.2 we report the performance of the system for the optimal parameter settings.

Although the use of all visual concepts available improves results on DS-2 dataset, the improvement is achieved for only a limited number of parameter settings and therefore cannot be considered robust enough. Moreover, in the case of DS-1 dataset improvement is not obtained for any parameter setting and/or choice of indicator. This finding supports our hypothesis that the concept selection is a critical step in our approach.

Table 2.2: MAP of our QES approach when all concepts and TF weights are used; statistically significant improvement over the baseline is indicated with "ˆ" (Wilcoxon Signed Rank test, $p = 0.05$)

Dataset	Concepts	Best Base.	QES	Oracle
DS-1	C-V374	0.2619	0.2363	0.3082
DS-2	C-V374	0.2621	0.268ˆ	0.3136
DS-2	MM 64	0.2621	0.2743ˆ	0.3136

Table 2.3: MAP of QES approach for DS-1 dataset when our concept selection approach on CU-VIREO 374 is used; statistically significant improvement over the baseline is indicated with "ˆ" (Wilcoxon Signed Rank test, $p = 0.05$)

Weights	Best Base.	QES	Oracle
TF	0.2619	0.2757ˆ	0.3082
TF-IDF	0.2619	0.2648	0.3082

2.7.2 QES applying the concept selection

In this section, we investigate the performance improvement that can be gained when applying concept selection. We first experiment with DS-1 and then analyze in Section 2.7.3 the capability of our approach to achieve a similar performance on the dataset DS-2 as well. For the DS-1 dataset from the entire CU-VIREO374 concept collection only 15 most informative concepts are selected.

The performance of our QES approach in this case is summarized in Table 2.3. It is still far from the ideal performance of the oracle, but it shows a moderate improvement over the best performing baseline, and also over the results in Table 2.2 (first row), where no concept selection was performed.

Robustness to parameter setting

To investigate the robustness of the retrieval performance in this case to parameter setting, we first investigate the QES performance for several values of threshold θ. In all cases, the number of top-N documents used to calculate the indicator is set to 20. This parameter value already yielded good results when calculating the coherence indicator on text vectors [85].

Table 2.4: MAP for different values of parameter θ; statistically significant improvement over the baseline is indicated with "ˆ" (Wilcoxon Signed Rank test, $p = 0.05$)

Weights	$\theta = 70\%$	$\theta = 80\%$	$\theta = 90\%$	Best Base.
TF	0.2735ˆ	0.2745ˆ	0.2757ˆ	0.2619
TF-IDF	0.225	0.2648	0.23	0.2619

The results are shown in Table 2.4. Normalized TF video representation appears to be more robust to parameter setting than TF-IDF, since it shows consistent improvement for various values of parameter θ. In [31], the suggested parameter value is 95%, but here it seems that the indicator calculated on concept-based features may be even more robust than the one calculated using conventional (text-based) TF or TF-IDF document representations.

Regarding the choice for x_{cv}, we investigate for which choice statistically significant improvements are obtained. In tables 2.2, 2.3 and 2.4 the statistically significant improvements over the baseline retrieval method are indicated with "ˆ". As a significance measure we adopt the Wilcoxon signed rank test ($p = 0.05$), commonly used in information retrieval. As indicated in the tables, almost all of the improvements obtained when the TF weights are used are statistically significant, which supports our conclusion that they are indeed more valuable for our purposes. The superior performance of TF weights might be a result of the fact that it is the pattern of concept occurrence (reflected in TF) rather than the absolute presence or absence of a concept in videos (encoded by IDF) that provides more helpful means of capturing semantic similarity. This effect may be specific to the distribution of concepts within video, since in text retrieval the IDF weight generally makes an important contribution. We conjecture that the IDF component is particularly sensitive to the noise of concept detectors and that a high IDF value for a particular concept might be caused by an erroneous detection. Finally, the reason for a lower performance might lay in the fact that we select a rather small subset of concepts that appear frequently in the collection, and thus the IDF component does not have a positive influence.

Table 2.5: MAP of QES approach for the DS-2 dataset when the concept selection approach on CU-VIREO374 concepts is used; statistically significant improvement over the baseline is indicated with "ˆ" (Wilcoxon Signed Rank test, $p = 0.05$)

Weights	Best Base.	QES	Oracle
TF	0.2621	0.2631	0.3136
TF-IDF	0.2621	0.256	0.3136

Optimality of the obtained results

Further we analyze whether our concept selection approach is capable of selecting the optimal threshold for the number of concepts to be used. Here we consider only TF weights, because, as shown in the previous section, they demonstrate a superior performance to TF-IDF weights. We gradually increase the number of top-N_c concepts in the lists produced based on frequency, variance and kurtosis criteria and thus the number of overlapping concepts. The best overall MAP of 0.2757 is obtained when 15 concepts are selected, which is the same result achieved with a concept set chosen using the automatically selected threshold. This finding confirms the capability of our approach to select the optimal value $\widetilde{N}_c$.

2.7.3 Generalization across datasets

For the DS-2 set, our concept selection approach extracts 32 representative concepts from the CU-VIREO374 concept collection. The performance comparison with the best performing baseline approach and the oracle indicator is shown in Table 2.5. Measuring the performance of the system for the varying number of selected concepts, as described in the previous section, reveals that in the case of TF-IDF weights our approach indeed selects the optimal number of concepts. In the case of TF weights, the maximal performance (MAP = 0.27) is obtained using the coherence indicator on 15 selected concepts (similar to the DS-1 set).

Similarly to the DS-1 set, when TF representation is used, a moderate improvement is achieved. TF-IDF representation again appears to be less robust and here it performs even worse than the best baseline approach (still outperforming the other three baselines). When the selection approach is applied to MediaMill concept collection, a subset of 14 representative concepts is selected. As an illustration, the automatically selected concepts are: *Building*, *Crowd*, *Face*, *Hand*, *Outdoor*, *Person*, *Per-*

Table 2.6: MAP of QES approach for the DS-2 dataset when the concept selection approach on MediaMill concepts is used; statistically significant improvement over the baseline is indicated with "ˆ" (Wilcoxon Signed Rank test, $p = 0.05$)

Weights	Best Base.	QES	Oracle
TF	0.2621	0.2688ˆ	0.3136
TF-IDF	0.2621	0.2673	0.3136

sonWalkingOrRunning, *Road*, *Sky*, *Street*, *TwoPeople*, *Urban*, *Vegetation* and *Waterscape*. The performance of QES approach using the Co indicator is shown in Table 2.6.

Both TF and TF-IDF concept-based feature variants yield a modest performance improvement and again the TF weights are performing slightly better, which is consistent with our previous findings.

2.7.4 Impact of quality of concept detectors

Compared to CU-VIREO374, the use of MediaMill concept set yields an increased robustness to parameter settings and as shown in tables 2.2, 2.5, 2.6 and 2.7, generally gives a higher performance improvement (in the terms of MAP). This is not unexpected, since the MediaMill system achieved the highest performance in e.g., TRECVID 2009 concept detection and interactive search tasks [98]. We can therefore conclude that the quality of concept detectors remains an important factor influencing the performance of our approach.

2.7.5 Alternative coherence indicators

In addition to the coherence indicator (Co) we test the usability of two alternative indicators of the topical clustering structure (mean-AIS and max-AIS). The experimental results show that when CU-VIREO374 concepts are used, the alternative indicators do not yield improvement on either DS-1 or the DS-2 set. However, when the MediaMill concepts are used on the DS-2 set the overall best performer for a wide range of parameter settings is the mean-AIS indicator. Table 2.7 summarizes the performance of our QES system on the DS-2 set, when the automatic concept selection approach is applied to the MediaMill concept set.

Wilcoxon signed rank test reveals that the obtained improvements are

Table 2.7: MAP of QES approach for the DS-2 dataset when the concept selection approach on MediaMill concepts and the mean-AIS indicator are applied; statistically significant improvement over the baseline is indicated with "ˆ" (Wilcoxon Signed Rank test, $p = 0.05$)

Weights	Indicator	Best Base.	QES	Oracle
TF	mean-AIS	0.2621	0.2719ˆ	0.3136
TF-IDF	mean-AIS	0.2621	0.27ˆ	0.3136

indeed statistically significant. The results from Table 2.7 are also consistent with our earlier findings. Namely, in our previous work [85] we showed that the mean-AIS and max-AIS indicators might be successfully used for query expansion selection when the videos are represented by the vectors of TF-IDF weights calculated on the automatic speech recognition transcripts text of the videos only. Moreover, the overall best-performing indicator in those experiments appeared to be mean-AIS. We conjecture that the performance improvement can be attributed to a higher sensitivity of mean-AIS indicator to the quality of concept detectors.

Mean-AIS indicator calculated on TF and TF-IDF weights gives consistent improvement in performance for different sizes of top-N video set over which it is calculated (i.e., $N = 5, 10, 15, 20$) when MediaMill concepts are used. This finding confirms that, depending on the quality of concept detector set, our approach is relatively robust to parameter setting.

2.7.6 Performance stability across queries

In this section we investigate how the improvement is distributed over queries. Our method predicts the correct query expansion in approximately 40% of time, but it also seems to make the correct prediction in the critical cases where the AP improves significantly. The indicator seems to make the error generally only in the cases when the coherence of the tops of different results lists is similar, but fortunately, the MAP values of these lists are also similar. For example, after the failure analysis of the results presented in Table 2.7, when the mean-AIS indicator is used on TF weights, we concluded that our indicator chooses the correct expansion in 43.9% of cases. Further analysis reveals that the indicator additionally chooses the second best expansion in 34.15% of queries and the errors were generally made in the case where the second best and the best results lists have very similar coherence of top results. Basically, our indicator selects the best or second best expansion in roughly 78% of queries. It is also important to note that our query expansion selection makes use of all available query expansions approaches. Namely, as shown in Table 2.1, the Google Sets expansion in general performs worse than the baseline retrieval, PRF and WordNet expansions, but there are queries for which it helps and our indicators seem to be capable of predicting such cases. For example, a failure analysis of the results presented in Table 2.6, when the Co indicator is used on TF weights, reveals that in the case of topical queries such as *dictatorship* and *youth programs*, the Google Sets expansion yields the best results and our indicator appears to be capable of detecting it. Further, in the case of queries *youth programs* and *patients* the best performing expansions are Google Sets and WordNet, while the generally better performing baseline and PRF expansion achieve 0 MAP. In both cases our indicator manages to select the best query expansion.

As explained in the introduction, we expect the performance of our approach to be influenced by several important factors, such as e.g., abstraction level of a particular semantic label, semantic and visual diversity of the videos relevant to that semantic label and the quality of visual concept detectors used. Tables 2.8 and 2.9 show for which semantic labels (queries) our query performance prediction approach succeeds in selecting an optimal expansion. Here, for both datasets and concept sets we use Co indicator on TF weights. A general observation can be made that the performance of our concept-based indicators is relatively independent of the abstraction level of a particular semantic label. In other words, the indicators manage to choose a correct results list for some more abstract (e.g., *daily life*, *politics*, *economy* and *history*) and some less abstract queries (e.g., *landscape* and *food*). We believe that in case of some abstract semantic themes our concept-based indicators are able to capture high-level stylistic similarities between videos, originating in television production rules. For example, political documentaries and talk shows usually feature several people talking about the subject. Further, in Table 2.9 we observe that for some semantic labels MediaMill concept detectors perform better, while in some other cases the better performing concept detector set is CU-VIREO374. This may be attributed to the fact that many concepts selected for those sets are different and not all concepts are equally representative of a particular semantic theme. Also, as shown in [35, 96], performance of concept detectors varies significantly within a concept de-

Table 2.8: Successfulness of our concept-based indicator, text-based alternative indicator and the combined indicator in predicting the optimal query expansion on the DS-1 set

Semantic Labels	CU-VIREO374	Text	Combo
work	-	-	-
poverty	-	+	+
biology	-	+	+
foreign workers	-	-	-
civil wars	-	-	-
computers	+	+	+
criminality	+	-	+
cultural identity	+	-	+
animals	-	-	-
zoos	-	-	-
economy	-	+	+
ethnic minorities	-	-	-
factories	-	-	-
families	-	-	-
celebrations	-	-	-
fraud	+	-	+
medicine	-	+	+
history	+	-	+
brain	+	-	+
assistance	-	-	-
journalists	+	+	+
children	-	-	-
landscapes	+	+	+
military personnel	+	-	+
musicians	-	-	-
music	-	-	-
physics	+	+	+
entrepreneurs	+	-	+
wars	-	-	-
seniors	+	+	+
press	-	+	+
politics	+	+	+
court hearings	-	-	-
elections	-	-	-
food	+	-	+
soccer	-	-	-
scientific research	+	-	+

Table 2.9: Successfulness of our concept-based indicators, text-based alternative indicator and the combined indicator in predicting the optimal query expansion on the DS-2 set

Semantic Labels	MediaMill	CU-VIREO374	Text	Combo
actors	+	-	-	-
work	-	+	+	+
asylum seekers	-	-	+	+
biology	-	-	-	-
books	-	+	-	-
criminality	-	-	-	-
daily life	+	+	+	+
dictatorship	+	+	+	+
animals	-	-	-	-
economy	+	+	-	+
ethnic minorities	-	+	-	-
factories	-	-	-	-
film	-	-	-	-
history	+	-	-	+
health science	-	-	-	-
brain	-	-	-	-
youth	+	+	+	+
youth programs	+	+	-	-
judicial system	-	-	-	-
children	+	+	-	+
artists	+	-	-	+
laboratories	-	-	-	-
agriculture	-	-	+	-
literature	-	-	-	-
people	-	-	+	-
military personnel	-	-	-	-
muslims	+	-	-	+
wars	-	-	-	-
seniors	-	-	-	-
patients	+	+	+	+
police	-	+	-	-
politics	+	-	+	+
paintings	-	+	+	+
writers	-	+	+	+
feature films	-	-	-	-
refugees	-	+	-	-
food	+	-	+	+
women	+	+	-	+
scientific research	-	-	-	-
housing	+	-	-	+
diseases	-	-	+	+

tector set, which further influences effectiveness and reliability of the set in capturing the semantic characteristics of a video. Finally, on the DS-2 set our concept-based indicators perform well for some semantically related queries, such as e.g., *children*, *youth* and *youth programs*. This observation supports our assumption that the correct decisions of concept-based indicators actually don't occur randomly, but depend on the quality of concept detectors and the degree to which a particular semantic theme is visually constraining.

2.8 Combined query performance indicator

While the experiments described in the previous section served to demonstrate that our concept-based video representation and the concept-based indicators of query performance are indeed promising solutions for semantic-theme-based video retrieval, here we compare their performance with the performance of text-based alternatives. As discussed in Section 2.3, recently proposed coherence-based indicators (e.g., [30], [31]), have been proven effective in a wide range of text information retrieval applications. In [85] we showed that post-retrieval coherence-based query performance indicators, such as those described in Section 2.5, might improve spoken content retrieval significantly. The retrieval framework used is similar to the one illustrated by Fig. 2.5, with, however, an important difference in video representation. For that, we exploit only the automatic speech recognition transcripts of the videos and represent each video as the vector of TF-IDF weights.

2.8.1 Text-based indicators on DS-1 and DS-2 datasets

In this experiment we use DS-1 set for exploring the parameter space and report results on DS-2 set. To simplify the analysis of indicator fusion, we limit the experiments to the coherence indicator Co only and report cases in which the other indicators perform better in terms of MAP or robustness to parameter setting. We choose to focus on the coherence indicator in this experiment also because it is the only indicator to yield performance improvement on both DS-1 and DS-2 sets when CU-VIREO374 concepts are used to represent videos. For text-based video representation, we index English translation of the automatic speech recognition transcripts and create vectors of TF-IDF weights. Preprocessing includes stemming and

Table 2.10: MAP of QES approach for DS-1 and DS-2 set when the coherence indicator Co is used with concept-based and text-based video representations; performance of indicator selection method is shown as well; statistically significant improvement over the baseline is indicated with “ˆ” (Wilcoxon Signed Rank test, $p = 0.05$)

Dataset	Best Base.	QES Concepts		QES Text		QES Combo		Oracle
		MAP	Corr.	MAP	Corr.	MAP	Corr.	
DS-1	0.2619	0.2757ˆ	40%	0.2624ˆ	30%	0.2846ˆ	54%	0.3082ˆ
DS-2	0.2621	0.2688ˆ	37%	0.2734ˆ	32%	0.2831ˆ	44%	0.3136ˆ

rigorous stopword removal, where each word appearing in more than N_s% of videos is considered to be a stopword. Our exploratory experiments show that the best results are obtained for $N_s = 20\%$. Further, similarly to [31], we experimentally prove that the text-based coherence indicator yields optimal performance when computed on top-5 documents using a high value for document similarity threshold $\theta = 95\%$. The performance on both datasets is reported in Table 2.10.

The best performer on DS-1 set is our proposed max-AIS indicator, scoring a MAP of 0.2648 when computed on top-5 documents. Generally, on both the DS-1 and DS-2 set max-AIS and mean-AIS appear to be more robust than the Co indicator, yielding improvement for a larger range of parameter settings. Finally, for the completeness of the analysis, we repeat the experiments representing videos as the vectors of normalized TF weights. Interestingly, normalized TF representation yields a similar performance improvement to TF-IDF for various values of stopword removal threshold $N_s \in [10\%, 90\%]$, which might suggest that some stylistic attributes of the conversational speech might be particularly useful for discriminating between videos based on the semantic theme.

2.8.2 Indicator selection

As discussed in Section 2.7.6, our best performing concept-based indicator, mean-AIS, on DS-2 set chooses a correct query expansion in roughly 40% of cases, while the first or second best expansion is selected in over 70% of cases. Therefore, we expect that fusion of text-based and concept-based indicators might lead to a further performance improvement.

To prove the concept, we choose to perform fusion through a simple voting strategy, acknowledging that a more sophisticated fusion approach might yield a higher performance improvement. First, we compute the indicators for the results lists generated in response to the original query and the three query expansions used and then select to use a more confident indicator for that query. We consider an indicator as more confident if it has a larger relative difference δCo^m between outputs for the most coherent and second most coherent results list.

$$\delta Co^m = \frac{Co_1^m - Co_2^m}{Co_2^m}, m \in \{c, t\} \tag{2.11}$$

In (2.11), Co_1^c and Co_2^c are the outputs of the concept-based indicator computed for the most coherent and second most coherent results list, while Co_1^t and Co_2^t are the corresponding outputs of the text-based indicator.

Table 2.10 shows the performance of our QES approach when: 1. Concept-based video representation (indicator) is used; 2. Text-based video representation (indicator) is used; 3. A more confident out of two computed indicators is selected. In a separate field, for each indicator we show a percentage of correctly selected expansions (i.e., percentage of cases in which the optimal query expansion is selected). As explained in the previous section, for the reasons of consistency and analysis simplification, we limit the experiment to coherence indicator Co only. On DS-1 dataset CU-VIREO374 concept set is used, while for DS-2 we make use of better performing MediaMill visual concept detectors.

The results indicate that selection of a more confident indicator brings additional performance improvement in terms of both MAP and ratio of correctly selected query expansions, which proves our starting assumption. The results presented in Table 2.10 indicate that the concept-based indicators yield a comparable performance to the state-of-the-art alternatives in the IR field, computed using the spoken content only. Furthermore, in the case of concept-based indicators performance improvement seems to be better distributed across queries (e.g., optimal results list/ query expansion is selected more often). An interesting observation can be made in Table 2.8: If either text-based or concept-based indicator manages to select the optimal results list, a combined indicator will succeed in the task as well. A similar, although not as constant, trend could be observed in Table 2.9, which further shows that even a simple combining of indicators can lead to a more reliable prediction. Finally, the experiments confirm our main assumption that the information relevant to a semantic theme can be extracted from the visual channel of the video and not only from its spoken content.

2.9 Discussion

We have presented an approach to semantic-theme-based video retrieval that uses shot-level outputs of visual concept detectors to automatically build video-level representations, here referred to as the Concept Vectors. These vectors are used to calculate coherence indicators that enable query expansion selection (QES) within a post-retrieval query performance prediction framework. The novel contribution of our approach is the effective combination of the output of automatic speech recognition and visual-concept detection, both known to be noisy, to achieve an overall improvement in retrieval of videos according to the semantic theme specified by the query. Our approach does not aim at obtaining hypothetical maximum performance on the given datasets, but rather to select the best out of available results lists for a given topical query in an unsupervised fashion. Concept Vectors are used to compare videos with each other instead of with the query. In this way, we are able to avoid any training that would be necessary to create a step that maps the query onto the appropriate concepts. Therefore, our approach has a potential to be used in a larger number of applications than the alternative solutions based on e.g., supervised learning.

A key advantage of our approach is its ability to make effective use of the noisy output of concept detectors. In fact, our Concept Vectors are designed to make optimal use of a given set of concepts, meaning that we do not necessarily need a guarantee that the set of concepts that we use provides a complete coverage of the semantic space of the collection. However, the starting concept set should provide a certain minimum required semantic coverage necessary for discriminating between videos at the level of a semantic theme. Also, given the concept detector sets of the same quality, the one providing a better semantic coverage is intuitively expected to yield a similar or better performance within our system.

Our experimental evaluation validated the effectiveness of our approach and confirmed that the automatic selection of concepts during the process of building the Concept Vector is critical for the retrieval performance improvement. Experiments also revealed that the automatically determined cut-off for the list of concepts to be used succeeds in approximating the optimal value. Further, it was shown that including the IDF factor provided no further performance gains, consistent with the conclusion that it is not so much the uniqueness of a concept in a video, but rather the frequency of that concept's appearance that best captures pair-wise similarity between videos in terms of semantic theme. The method for automatic selection of concepts to be used to build the Concept Vector was shown to be transferable in an unproblematic manner to an unseen dataset of a similar type. Changing datasets does, however, require a re-optimization of the parameters involved in calculating the coherence indicator, namely the θ cutoff and also the number of top-N documents used.

The improvement yielded by the approach is distributed relatively well across the board, i.e., its benefit is not localized to only certain types of queries. In particular, there is no apparent correlation between the absolute number of documents relevant to a particular query within the collection and the effectiveness of our QES approach. This observation supports our claim that the applicability of our approach generalizes well across different kinds of queries presented to the system, and in particular to new queries with new properties. The efficacy of the approach was shown to have a sensitivity to the quality of the concept detectors, with better performing concept detectors yielding higher improvement of QES.

The automatic approach for generating Concept Vectors involves a relatively small number of concepts. If a small set of well-chosen concept detectors in certain scenarios such as the one described in this chapter is sufficient to improve the results of semantic-theme-based retrieval, a productive avenue for the development of concept detectors is to concentrate on achieving high quality for a small number of detectors and not on training concept detectors that will cover the entire conceivable semantic space.

We demonstrate that not only spoken content of the video, but also information extracted from the visual channel can be successfully exploited for discriminating between videos based on the semantic theme. Finally, here we show that a simple combination of "unimodal" coherence indicators of query performance, exploiting text-based and concept-based video representations, might further improve retrieval performance. More specifically, for each query we first compute text-based and concept-based query performance indicators and then automatically select the more confident indicator to obtain a higher performance improvement, both in terms of overall MAP and percentage of correctly selected expansions. Experiments reveal that e.g., our combined query performance indicator makes a correct decision for 30% queries more than a state of the art text-based alternative.

Our future work will involve investigation into the further refinement of the approach to building concept-based video-level representations. In particular, we are interested in exploiting not only the frequency of occurrence of concepts, but rather detailed information about their occurrence patterns, including distributional properties such as burstiness and also co-occurrence with other concepts. Finally, we are interested in investigating methods for automatically estimating the optimal parameter settings for QES, in determining the lower bound of concept detection quality necessary for a concept detector to be useful in our method and also determining the exact nature of the collection-specific properties that make our approach more or less suitable for a particular retrieval task.

Chapter 3

Visual summarization of geographic areas

While the previous chapter considered a use-case of an unlabeled video collection, in this chapter we move to information-rich social media. We present here a novel approach to visual summarization of geographic areas using community contributed images. The approach, which jointly utilizes visual content of the images, user-generated annotations as well as the information about users and their social network, aims at illustrating all relevant aspects of a geographic area by selecting the most representative images for each aspect. In addition, a novel evaluation protocol is proposed, which utilizes information associated with the images only and does not require judgment of human evaluators.

This chapter is published as: Stevan Rudinac, Alan Hanjalic, Martha Larson. Generating visual summaries of geographic areas using community contributed images. *IEEE Transactions on Multimedia*, 2013, Early Access Article.

3.1 Introduction

Availability of affordable image and video capturing devices as well as rapid development of social networking (e.g., Facebook, MySpace) and content sharing websites (e.g., Flickr and YouTube) has led to the creation of a new type of content, popularly known as social media. In such environments, multimedia content (images, video, music) is usually accompanied by user-generated metadata, such as title, description, tags and comments. While these types of metadata can be referred to as *explicit* ones, *implicit* metadata can be derived as well, like for instance those containing the information on the uploader and user relations inferred from users' interactions with the images and their activity in a social network related to these images.

In this chapter we present an approach to automatic creation of visual summaries of geographic areas using community-contributed images and related explicit and implicit metadata. The goal is to produce a visual summary of the (e.g., circular) area surrounding a given location, e.g., a landmark, hotel or a museum, where the location is specified by its geo-coordinates (geotags). The summary should be as informative about the location, but at the same time as compact as possible. The approach is motivated by the assumption that a person deciding on whether to visit a particular location may be guided to a large extent by his impression about the surroundings of that location (e.g., when choosing a hotel). Compact and informative visual summaries could help improve time efficiency and effectiveness in interacting with typical interfaces for location recommendation and visualization (e.g., hotel reservation websites) and in using interactive map exploration tools to generate an impression about the location. Two examples of visual summaries for the area around a location at the Champs-Élysées avenue in Paris, France, are shown in Fig. 3.1. Since summary (a) shows images of a single dominant landmark, we target a summary of the type (b), which includes various aspects of the area, including the *mainstream* ones like popular landmarks (L'arc de triomphe or the historical George V hotel), but also the *non-mainstream* ones, such as e.g., exclusive stores typical for Champs-Élysées and its surroundings. In the text that follows we consider the locations, objects and events to be mainstream if they have been found interesting by many users and therefore appear often in the image collection. Similarly, non-mainstream or *off the beaten track* locations, objects and events are those that are found interesting by a smaller group of users and therefore appear less frequently in the collection.

In our approach, we choose to integrate the available social media resources using a graph-based model and to steer the process of visual summary creation by the information derived from the analysis of the model. Although the topology of our graph is similar to the one used in other related work, as e.g., [8, 13, 68, 105], we propose a novel method for automatic weighting of the graph edges in order to reflect the contribution of each modality, namely text, visual features and user relations, to the overall performance of our visual summarization algorithm. Since the analysis of methods for computation of affinities (similarities) between nodes in the graph is not the main focus of this chapter, we compute them using a well-known state-of-the-art method. Assuming that the informativeness of the visual summary is determined by representativeness and diversity of the images it consists of, we further propose a novel method that utilizes the computed multi-modal node similarities to first evaluate the representativeness of each image in the graph and then to jointly enforce representativeness and diversity in the target image set. Finally, we also propose a protocol designed to evaluate the quality of the generated image set that does not require input of human annotators, but rather exploits available metadata associated with the images.

In Section 3.2 we first address related previous work and then present the rationale behind our approach in Section 3.3. The proposed algorithm for generating visual summaries is described in Section 3.4, while the new evaluation protocol is introduced in Section 3.5. This protocol was deployed in the experimental setup described in Section 3.6 to produce experimental results that are presented and analyzed in Section 3.7. The discussion in Section 3.8 concludes the chapter.

3.2 Related work

While diversified image search results are preferred [10, 90], it was also found that the users are more sensitive to irrelevant than to duplicate images in the results list [41]. Therefore, increasing diversity without relevance deterioration poses a major challenge in the area [72]. State-of-the-art image search (set) diversification approaches can be divided into several categories according to whether only visual content, text associated with the images, automatically generated metadata, users' activity statistics or a combination of these resources is exploited.

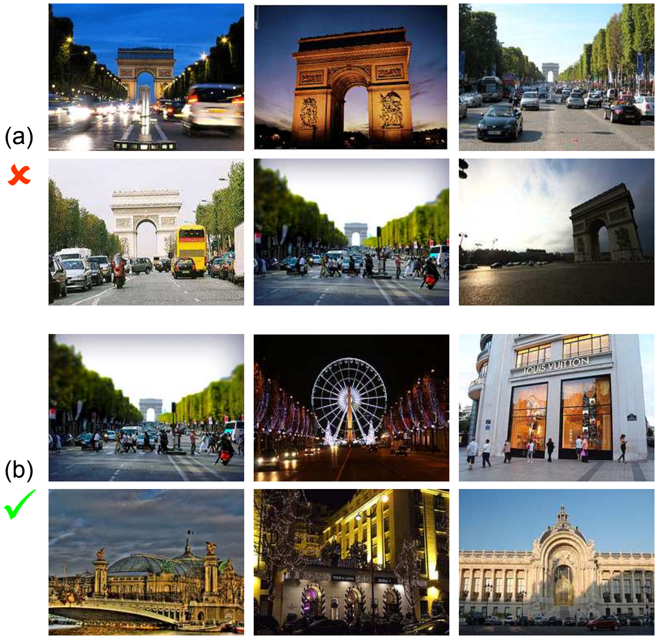


Figure 3.1: Two examples of visual summaries of the area around a location at the Champs-Élysées Avenue in Paris, France; a) example showing images of a dominant landmark only; b) a better example showing various aspects of the area - famous landmarks, stores, hotels etc. All images are downloaded from Flickr under CC license.

Perhaps the most intuitive approach to diversifying image search results is to perform image clustering in visual domain and then select a representative of each cluster to be presented to the user [106]. We anticipate, however, that in the use scenario envisioned in this chapter the expected high diversity of the visual content in the images taken in an arbitrary geographical area will make the clustering task a challenging one, especially if this content is not dominated by distinct landmarks or other prevailing objects and scene aspects.

Following a different approach in [99], Song et al. propose two scores, the topical richness score, which measures the information content contributed by each image added to the results list, and the diversity score, which measures the topical coverage of the (final) image results list. The images tagged with scarce topics (topics that are rarely included in the image set) are favored over rich topics widely distributed throughout the collection. Note that the authors consider each word used to tag images to be a separate topic. To calculate topical richness, first the topical similarities between images in the candidate set, output of a particular image search engine, are calculated. The topical richness of a given image is further calculated iteratively, similar to the PageRank [67] algorithm, by aggregating the topical richness of its neighbor images. The contribution of each image in the final score is proportional to its similarity to the image whose topical richness is being calculated. While the proposed algorithm performed well on a dataset of 20.000 illustrations tagged with a total of 718 unique words, this number of annotation words is rather small compared to the social media context where images are annotated (in the form of titles, tags and comments) with thousands or even tens of thousands of different words. Furthermore, Song et al. perform the image diversification based on the associated text only, while the visual content of the images is not taken into account. This choice could be considered suboptimal in a general case in view of the results of the ImageCLEF Photo Task 2009, which indicate that the best performing approaches in terms of precision and diversity exploit both the visual content of the images and the associated text [72].

Recently, the ImageCLEF Photo Task focused on diversification of image search results. The dataset used in the 2009 edition of the benchmark was composed of 498.920 images from a Belgian press agency. To promote diversity, *cluster recall* was used, which basically measures the ratio of retrieved clusters in the top K results of the results list and the total number of possible clusters associated with a given search topic. In order to balance the diversification effort, precision at 10 (P@10) was used as another performance measure. A total of 50 topics were used for evaluation and each of them was associated with a certain number of clusters. For example, relevant clusters associated with the topic *Clinton* were *Hillary Clinton*, *Obama Clinton* and *Bill Clinton*. This is, to the best of our knowledge, the first standardized dataset used for evaluating image search diversification approaches. However, it is not applicable in the context of this chapter for several reasons. In our approach, we would like to exploit user contributed and annotated images available in e.g., content sharing websites, such as Flickr. The type and quality of information in the social media context is often very different from the professional content. Our visual summaries target visual presentation of geographic area within a set radius (e.g., 1km from a given location) and the images captured there might depict a large number of points of interests, so the tag sets of relevant images might not even intersect. Some queries used at ImageCLEF Photo Task 2009 have a geographic connotation, but they are inappropriate for our task since they often either refer to larger geographical units or include clusters referring to organizations and events with a location prefix. For this reason, we collect our own data set to develop and evaluate our algorithm.

In [41], Kennedy and Naaman present a multimodal approach to selecting diverse and representative image search results for landmarks. They also rely on both the visual information in the images and the user-contributed tags for these images. However, contrary to our objective to show the surroundings of a given (e.g., landmark) location, they focus on selecting the best views of the landmark itself. The system demonstrated good performance on a dataset of 110.000 Flickr images of the San Francisco area.

In [75] Popescu et al. propose an approach that leverages Flickr images and the associated metadata for discovery and analysis of a large number of tourist trips and for the recommendation of new one-day trips. Although the approach does not explicitly address the problem of visual summarization of a given geographic area, we mention it here because of a common general application domain and the fact that they also rely on community-contributed content. In other work related to location recommendation for tourists, Cao et al. [9] start from a large number of Flickr images and cluster them using the associated geo-coordinates. In the following step, the most frequent tags as well as the visually representative images are selected to represent each geo-cluster. Although we find the general aim of this approach to be related, it differs from our approach significantly. Namely, our visual summarization approach does not make use of the geo-coordinates and we deploy them only in our proposed evaluation protocol, described in detail in Section 3.5. Also, to select representative images for a geo-cluster the approach presented in [9] utilizes visual features only, which may be a suboptimal choice [72]. In terms of its general objective, our approach is closely related to [24], where the authors target a summarization of touristic destinations by mining user-generated travelogues. In [69], the authors pursue a similar idea by first mining the online travelogues to detect representative location-specific tags. These tags are further used to select relevant and representative images amongst the location-relevant images from Flickr. Finally, a particular location is characterized by both the travelogue text information and representative images. While the authors evaluate their approach using 20 touristic cities as destinations, it is unclear whether the approach would be as efficient for summarizing smaller geographic areas that are often not covered by high quality (or any) travelogues.

3.3 Approach rationale

In this section we discuss in more detail the rationale underlying the proposed approach for creating a visual summary of a geographical area.

As illustrated in Fig. 3.2, the input into our visual summarization algorithm is a location, e.g., a hotel, landmark or a museum, which is specified by its geotag. We then select all images available on Flickr that are likely to have been taken in the surroundings of that location. For the practical implementation of our algorithm we constrained this selection to the range of the radius of 1km around the input location and selected the initial set of images on the basis of their geotags. We note here, however, that the availability of geotags is not critical for this starting step of our approach. Social media can be efficiently filtered even when the geotags are not available, e.g., by using event-based and area/landmark-specific tags. For example, using event-based tags, such as e.g., "*ACM Multimedia 2010, Florence*", it would be possible to isolate a large number of user-contributed Florence images and particularly those captured in the vicinity of the conference venue. Often, events create their own pages in the social networking and content sharing websites which makes the task even easier and more realistic. Furthermore, in [41] Kennedy and Naaman discover a large number of landmark images in Flickr by searching for landmark-specific tags. They also show that a large number of images captured in the vicinity of landmark are tagged with the landmark-specific tag, although they did not actually depict the landmark itself. In their approach they attempt to eliminate those images and keep the representative views of the landmark only, but for our task all images they initially collected would be valuable.

Inspired by successful existing graph-based approaches to resolving problems in the domain of social media retrieval [8, 13, 105], we also choose a graph-based model as the basis for our summarization algorithm. To model relations between images captured within a certain radius from a given geographic location as well as the associated explicit and implicit metadata, we construct a multi-modal graph consisting of several types of nodes and edges.

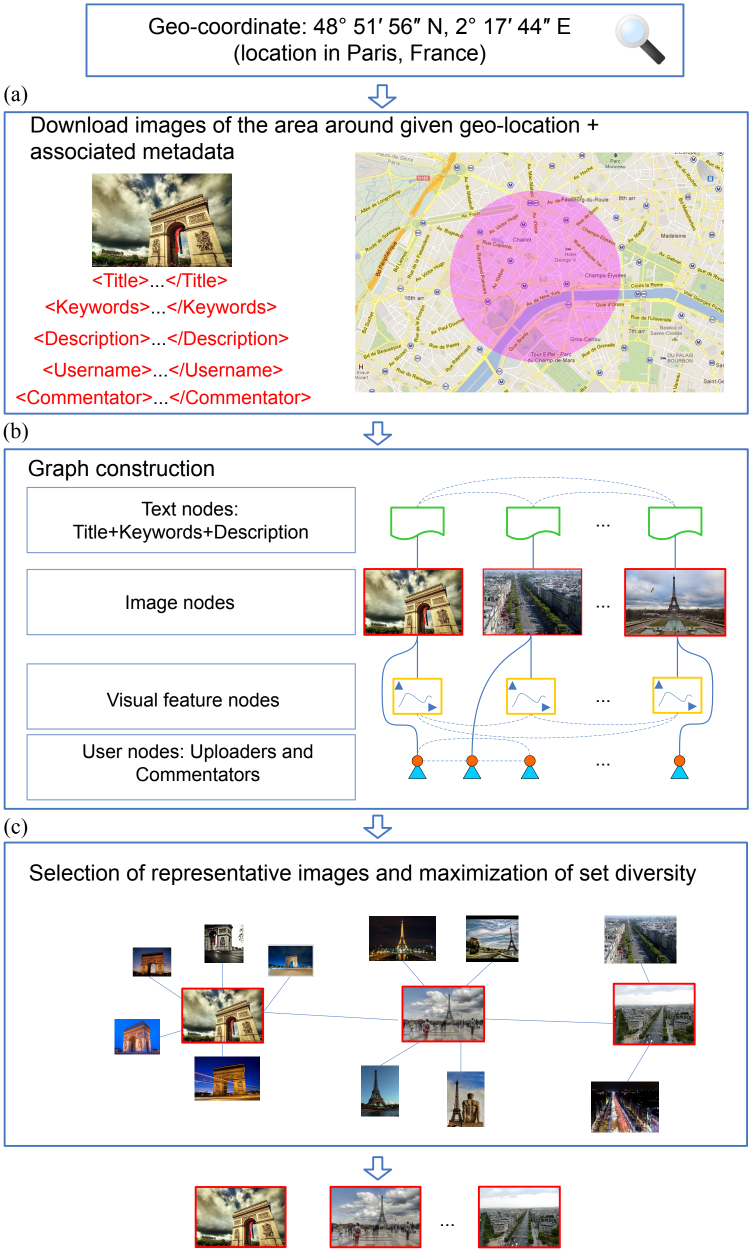

Geo-coordinate: 48° 51′ 56″ N, 2° 17′ 44″ E
(location in Paris, France)
(a)
Download images of the area around given geo-location + associated metadata
<Title>...</Title>
<Keywords>...</Keywords>
<Description>...</Description>
<Username>...</Username>
<Commentator>...</Commentator>
(b)
Graph construction
Text nodes: Title+Keywords+Description
Image nodes
Visual feature nodes
User nodes: Uploaders and Commentators
...
(c)
Selection of representative images and maximization of set diversity
...

Figure 3.2: Illustration of our approach to visual location summarization. The approach consists of three main steps: (a) collecting the initial image set and related metadata, (b) multi-modal graph construction and (c) using the graph to filter the initial image set for representative and diverse images.

In view of the fact that the initial step of the approach does not depend on the availability of geotags in community-contributed image collections, we choose to also keep our summarization algorithm as generic as possible and do not rely on geotags as an information resource when designing our graph model. Instead, we rely on the visual content of images, user-generated annotations and user relations derived from users' interaction with the images and their activity in a social networks related to the images, as described in Section 3.1. Although user-generated annotations are often noisy and imprecise [40], they show a high potential for improving the performance of modern image and video search engines. Similarly, a substantial amount of recent work, and in particular those addressing the collaborative filtering problem, have brought to light the high usefulness of the information related to users' activity in a social network for realizing various applications in the social media domain (e.g., [43]).

Although popularity indicators, like the image view count, could also be considered generally as a useful information resource to be included in the abovementioned graph model, we do not rely on such indicators when designing our summarization algorithm. While analyzing the notion of "popularity" in general goes beyond the scope of this chapter, we can state that a popularity indicator in the specific case of visual summary creation has a similar negative effect as in general collaborative recommendation systems [73]. It is namely likely to bias the summary towards the images showing the "mainstream" scenes and objects (e.g., "Eiffel Tower") and at the same time create a long tail of practically inaccessible, but potentially valuable "off-the-beaten-track" images. Although view count might be implicitly taken into account via user connections in the social subgraph (cf. Fig. 3.3), we noticed that the benefits of exploiting user connectivity are stronger than the potential flaws due to the mainstream bias.

3.4 Summarization algorithm

In this section, we describe three main algorithmic modules that are related to steps (b) and (c) in Fig. 3.2.

3.4.1 Graph construction

Let $G = (V, E)$ be our undirected graph with the set of nodes V and the set of edges E. Here we choose to use the undirected graph because, as will be described below, the relations between nodes (e.g., visual, text and user similarities) are symmetric. The graph has four layers and is illustrated by the structure given in Fig. 3.3.

Nodes

The graph consists of the following sets of nodes:

- *Image nodes* $I = \{i_1, i_2, \ldots, i_N\}$: For each of N images of a particular location an image node is introduced.
- *Visual feature nodes* $F = \{f_1, f_2, \ldots, f_N\}$: For each of N images in the initial set a visual feature node is added. Here, we represent the visual content of an image with a vector composed of two low-level feature components - Gabor texture features and local color moments extracted over the 5×5 regular rectangular grid. Adding additional visual features is straightforward, requiring for each new feature only the introduction of a layer with N nodes. Depending on the matching strategy and the similarity metrics used, it may also prove sensible to expand the feature vectors in the existing visual feature nodes with new feature components.
- *Term nodes* $T = \{t_1, t_2, \ldots, t_N\}$: A term node is added for each image in the set. To index images we exploit user-generated title, description and tags. We opt not to weight those fields individually, but rather consider all the text associated with an image to be a single document. Finally, each image is represented by a vector of TF-IDF (term frequency - inverse document frequency) weights [89].
- *User nodes* $U = \{u_1, u_2, \ldots, u_{N_u}\}$: A node is added for each of N_u users uploading an image related to a given location or commenting somebody else's image.

The final set of nodes V in our graph G is equal to:

$$V = I \cup F \cup T \cup U \tag{3.1}$$

Edges

Similar to [68] and [105], we introduce two types of edges in our graph, namely the attribute and similarity edges.

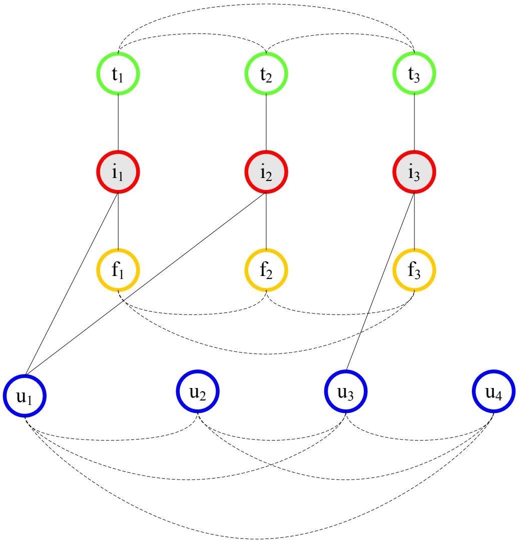


Figure 3.3: Illustration of the proposed four-layer graph structure.

- *Attribute edges*: These edges are marked with solid lines in Fig. 3.3. An attribute edge is added between an image and each of its attributes - visual feature, author (user who captured/uploaded the image) and text nodes. Note that a single user might upload multiple images, while some users don't have any uploads and therefore their nodes are not linked to any image.

- *Similarity edges*: These edges are marked with dashed lines in Fig. 3.3. This set of edges link nodes of the same type.

The edges linking visual feature nodes are weighted by the visual similarity score, computed using a Gaussian kernel, that is

$$\mathbf{W}_f(l,j) = sim(\mathbf{f}_l, \mathbf{f}_j) = \exp\left(-\frac{\|\mathbf{f}_l - \mathbf{f}_j\|^2}{2\sigma^2}\right) \tag{3.2}$$

where $\mathbf{W}_f$ is the $N \times N$ matrix of weights.

To weight the edges between user nodes we compute implicit user similarity based on how many images are related to both users. An image is related to both users if e.g., one of them uploaded the image and the other one commented on it and/or if both users commented on that image. The corresponding user similarity measure is computed as

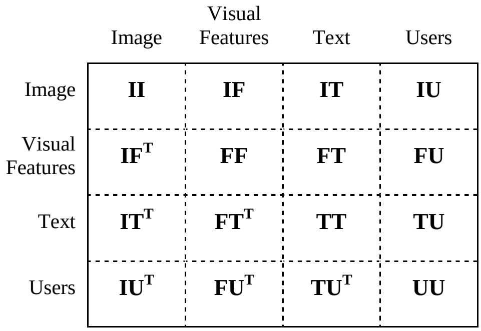


	Image	Visual Features	Text	Users
Image	**II**	**IF**	**IT**	**IU**
Visual Features	$\mathbf{IF}^{\mathrm{T}}$	**FF**	**FT**	**FU**
Text	$\mathbf{IT}^{\mathrm{T}}$	$\mathbf{FT}^{\mathrm{T}}$	**TT**	**TU**
Users	$\mathbf{IU}^{\mathrm{T}}$	$\mathbf{FU}^{\mathrm{T}}$	$\mathbf{TU}^{\mathrm{T}}$	**UU**

Figure 3.4: The adjacency matrix A of our graph G

$$\mathbf{W}_u(l,j) = sim(u_l,u_j) = \frac{|I'_l \cap I'_j|}{|I'_l \cup I'_j|} \tag{3.3}$$

where $I'_l, I'_j \subset I$ are the sets of images uploaded/ commented by the users u_l and u_j and $\mathbf{W}_u$ is the $N_u \times N_u$ matrix of user similarities. To obtain a better approximation of user links we consider images from all available locations and not only the one we are constructing the graph for.

The edges linking text nodes are weighted using the cosine similarity between the vectors of TF-IDF weights:

$$\mathbf{W}_t(l,j) = sim(\mathbf{t}_l,\mathbf{t}_j) = \frac{\mathbf{t}_l \times \mathbf{t}_j}{||\mathbf{t}_l||\,||\mathbf{t}_j||} \tag{3.4}$$

where $\mathbf{W}_t$ is the $N \times N$ matrix of weights. All similarity values are scaled to fit the $[0,1]$ range.

Additional edge weights

We see edge weighting in the service of an effective multi-modal fusion as a major challenge in the graph-based algorithms for multimedia indexing, retrieval and summarization. The majority of recent work deploying graphs for these applications either discards the analysis of an individual modality's contribution and only ensures that all similarities are scaled to the same range [8], or experimentally assigns fixed weights to different modalities [105]. However, Clements et al. [13] pointed out how critical modality weighting is in the applications resembling the one addressed in this chapter.

To analyze the contribution of each individual modality to the performance of our visual summarization algorithm, we construct three subgraphs of the graph G from Fig. 3.3, each consisting of only image nodes and either feature, text or user nodes, and use them to replace the graph G in the summarization process. Let the vertex sets of these subgraphs be $V_f' = I \cup F$, $V_t' = I \cup T$ and $V_u' = I \cup U$, respectively. If the performance of our summarization system for a particular location in these individual cases is given with π_f, π_t and π_u (cf. Section 3.7.3), we define modality-dependent weights β_f, β_t and β_u for that location as $\beta_f = \frac{\pi_f}{\pi_f+\pi_t+\pi_u}$, $\beta_t = \frac{\pi_t}{\pi_f+\pi_t+\pi_u}$, $\beta_u = \frac{\pi_u}{\pi_f+\pi_t+\pi_u}$. We use these weights to modify both the attribute and similarity edges, such that the importance of each modality is properly reflected. Starting from the edge weights $\mathbf{W}_m$ assigned as described in the previous section, the final edge weights $\mathbf{W}_m'$ are given as: $\mathbf{W}_m' = \beta_m \mathbf{W}_m, m \in \{f, t, u\}$. Note that we compute modality-dependent weights β_f, β_t and β_u for each location independently, which improves robustness and scalability of the approach.

Adjacency matrix of the graph

The adjacency matrix $\mathbf{A}$ of our graph G is illustrated in Fig. 3.4. It consists of the following submatrices:

- $\mathbf{II}_{N\times N} = [0]_{N\times N}$: Since the multi-modal similarities between image nodes $i_l, l = 1, \ldots, N$ are not known, $\mathbf{II}$ matrix is filled with zeros.

- $\mathbf{IF}_{N\times N} = \beta_f \mathbf{I}_N$: Weights on attribute edges linking image nodes i_l, with the corresponding visual feature nodes $f_l, l = 1, ..., N$ are multiplied with the overall modality-dependent weight β_f. Note that $\mathbf{I}_N$ is an $N \times N$ identity matrix.

- $\mathbf{IT}_{N\times N} = \beta_t \mathbf{I}_N$: Weights on attribute edges linking image nodes i_l, with the corresponding text nodes $t_l, l = 1, \ldots, N$ are multiplied with the overall modality-dependent weight β_t.

- $\mathbf{IU}_{N\times N_u}(l,j) = \left\{ \begin{array}{ll} \beta_u, & u_j = \mathit{uploader}(i_l) \\ 0, & \text{otherwise} \end{array} \right.$: Attribute edges connecting image nodes $i_l, l = 1, \ldots, N$ with their uploaders' nodes $u_j, j = 1, \ldots, N_u$ are assigned the overall modality-dependent weight β_u. Remember that the total number of users in our system is N_u and as users we consider both uploaders (authors) and commentators (users commenting an image).

- $\mathbf{FF}_{N\times N} = \beta_f \mathbf{W}_f$: Weights of the edges linking visual feature nodes.
- $\mathbf{FT}_{N\times N} = [0]_{N\times N}$: Since the multimodal similarities between visual feature nodes and the text nodes are not known, we fill the corresponding matrix with zeros.
- $\mathbf{FU}_{N\times N_u} = [0]_{N\times N_u}$: Since the multimodal similarities between visual feature nodes and the user nodes are not known, we fill the corresponding matrix with zeros.
- $\mathbf{TT}_{N\times N} = \beta_t \mathbf{W}_t$: Weights of the edges linking text nodes.
- $\mathbf{TU}_{N\times N_u} = [0]_{N\times N_u}$: As the multimodal similarities between text nodes and the user nodes are not known, we fill the corresponding matrix with zeros.
- $\mathbf{UU}_{N_u\times N_u} = \beta_u \mathbf{W}_u$: Weights of the edges linking user nodes.

The adjacency matrix A is column-normalized such that the values in each column sum to 1.

3.4.2 Selection of representative images

In order to select representative images for a given location, multi-modal affinities (similarities) between items in the graph need to be computed first. A plethora of methods can be used for this purpose [115]. Since the analysis and comparison of those methods is not the focus of this chapter, we opt to utilize *random walk with restarts* (RWR) over the graph described above. RWR is a well-known concept in the field of information retrieval, with one of its best known application being the Google PageRank algorithm [67]. It has also been successfully used in image retrieval and tagging [68, 105], as well as for collaborative recommendation [43]. Again, this particular design choice does not affect the generality of our method since any other state-of-the-art approach for computing the similarity between items in the graph could be used instead.

Our algorithm for selection of representative images to visualize a given geographical area can be outlined as follows. First, we initiate the RWR from each image node in the graph, one at a time. In each step, a random walker selects randomly an outgoing edge or decides to restart the random walk with probability α. In the initial work of Page et al. [67] $\alpha = 0.15$ was reported as the optimal value, but the later works using RWR in image tagging and retrieval applications [68, 105] found the optimal value to be significantly higher. Therefore, throughout this chapter we use $\alpha = 0.5$. The stationary probabilities $\mathbf{p}$ of all nodes in the graph are obtained after solving the equation

$$\mathbf{p} = (1 - \alpha)\mathbf{A}\mathbf{p} + \alpha\mathbf{v} \tag{3.5}$$

where $\mathbf{A}$ is the $|V| \times |V|$ adjacency matrix of the graph G. The random walk is initialized using a $|V| \times 1$ restart vector $\mathbf{v}$. All values in the restart vector are set to 0, apart from the position of the starting image node, which is set to 1. The direct algebraic solution of (3.5) is given as

$$\mathbf{p} = \alpha(\mathbf{I} - (1 - \alpha)\mathbf{A})^{-1}\mathbf{v} \tag{3.6}$$

In the case of large graphs where the matrix inversion is computationally intensive or practically infeasible, (3.5) is efficiently solved in an iterative manner. We repeat RWR for each image node i_l in the graph G by setting the l-th position in the restart vector $\mathbf{v}$ to 1 and store stationary probabilities of image nodes (i.e., first N values from $\mathbf{p}$) in the l-th column of the matrix $\mathbf{S} = [s_{lj}]_{N\times N}$. For every pair of images $\{i_l, i_j\} \in I$, s_{lj} represents a multimodal similarity between them.

We conjecture that the representative images should be salient, or in other words, similar to many other images captured in the vicinity of a given location. Therefore, for an arbitrary image i_l we compute the sum of its similarities to all other images in the graph G as

$$q_l = \sum_{j=1:N, j\neq l} s_{lj} \tag{3.7}$$

When calculating $\mathbf{q} = [q_1 \; q_2 \; \dots \; q_N]$ we don't take into account image self-similarities, because relatively high self-similarity values would enable visual outliers to negatively affect the results.

In the following step we sort the images according to the increasing q value and define the representativeness score RS for each image to be equal to the rank position of the image in the sorted list (i.e., image with the smallest q will have $RS = 1$ and image with the highest q will be assigned $RS = N$). As the first element in the target image set (further referred to as the *optimal set*, OS), we select the image with the highest representativeness score. The outcome of sorting images according to their representativeness and selection of the first image for OS is illustrated in STEP 1 of Fig. 3.5.

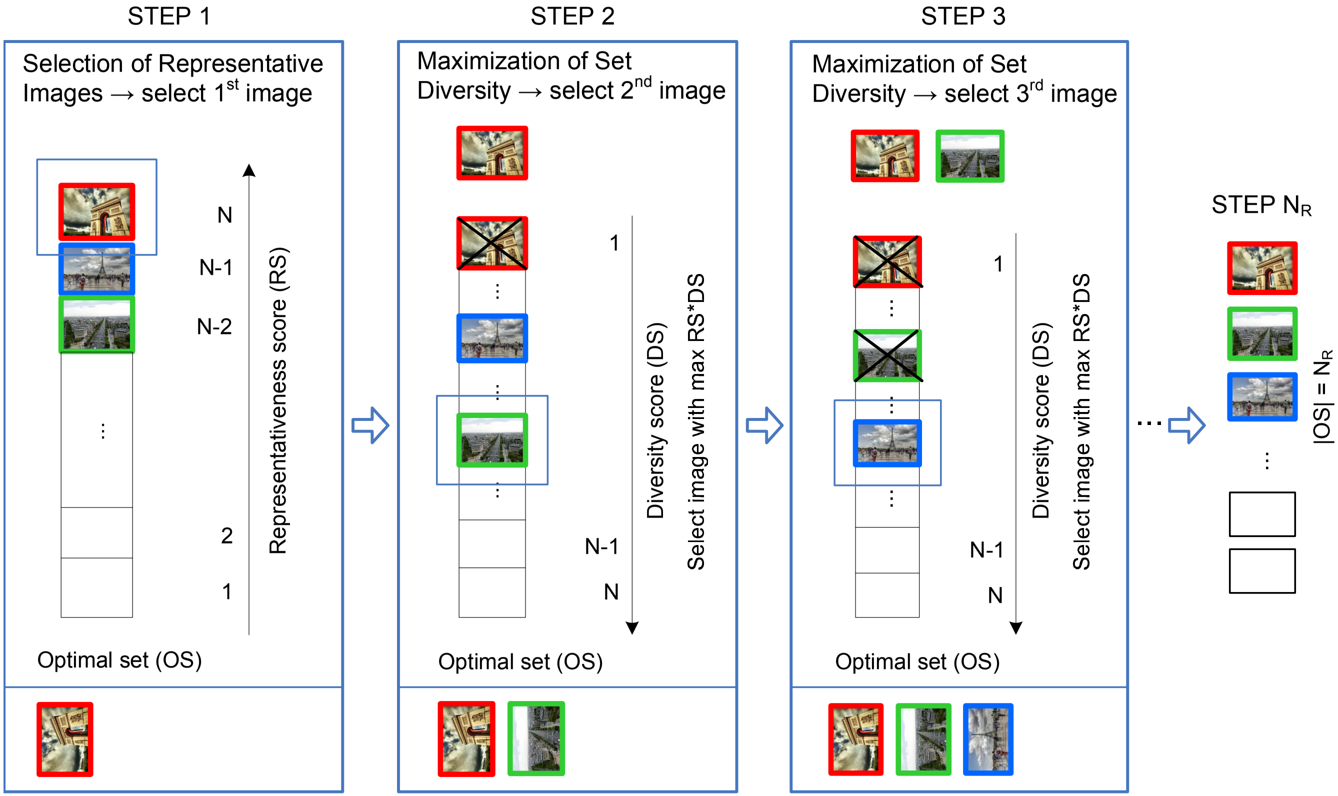

STEP 1
Selection of Representative Images → select 1st image
N
N-1
N-2
2
1
Representativeness score (RS)
Optimal set (OS)
STEP 2
Maximization of Set Diversity → select 2nd image
1
N-1
N
Diversity score (DS)
Select image with max RS*DS
Optimal set (OS)
STEP 3
Maximization of Set Diversity → select 3rd image
1
N-1
N
Diversity score (DS)
Select image with max RS*DS
Optimal set (OS)
STEP N_R
|OS| = N_R

Figure 3.5: Illustration of our approach for selection of diverse and representative images.

3.4.3 Maximization of set diversity

After the most representative image is selected, we select the next images for OS in the iterative fashion, as shown in the STEPS 2 - N_R of Fig. 3.5. The key insight that we exploit in order to make the resulting set of images OS as diverse and representative as possible is the following: To enforce both representativeness and diversity, the next image selected for the OS should be as dissimilar as possible with the previously selected image(s) and have at the same time a high representativeness score. Therefore, we initialize the RWR setting values in the restart vector $\mathbf{v}$ to $1/|OS|$ in the position of already selected images and 0 otherwise. Stationary probabilities of all nodes in the graph are computed using (3.6) and the first N values from $\mathbf{p}$ are stored in $\mathbf{p^{(OS)}} = [p_1^{(OS)}\ p_2^{(OS)} \dots p_N^{(OS)}]$. The stationary probabilities $p_j^{(OS)}, 1 \leq j \leq N$ reflect the similarity between each image in the graph and the images from OS. Further we sort the elements of $\mathbf{p^{(OS)}}$ in the decreasing order. For each image, we define its diversity score DS as being equal to its position in the sorted list, such that the image most similar to already selected images in the OS has $DS = 1$, and the image that is least similar to the images in the OS has $DS = N$. Finally, we select the image with the highest $RS * DS$ value amongst the images that have not already been included in the OS. This procedure for image selection is repeated until the desired number N_R of representative and diverse images ($1 \leq N_R \leq N$) have been selected.

3.5 Evaluating visual summaries

3.5.1 Rationale

Proper evaluation of social media retrieval algorithms often requires extensive user tests. Until recently, due to the difficulties in finding proper subjects willing to perform such tasks and due to the evaluation costs, the number of evaluators was often relatively small [99, 106]. Rapid development of Web 2.0 technologies has made crowdsourcing a popular and efficient tool for tasks such as annotation and evaluation. Services like Amazon Mechanical Turk allow researchers to perform large-scale user tests at a reasonable cost [101]. However, since algorithmic development usually requires intensive experimentation, it would be neither practical nor cost-efficient to evaluate each parameter setting or algorithm variant in this way.

In view of the above, we investigated the possibilities to test the quality of the generated visual summary of a given geographic area based on easily accessible objective information. The evaluation protocol we propose in this chapter does not require input of human annotators, but rather exploits specific metadata associated with the images, such as their geotags. The evaluation measure is intended to reflect the conditions on image sets that must necessarily be fulfilled in order for users to find them representative and diverse. As such the measure is useful for our stated purpose of cost-efficient algorithm development. Our evaluation protocol is inspired by the recent user preference studies on spatial diversity in image retrieval such as e.g., [101], which prove that users have a strong preference towards spatially diversified results.

We foresee that the applicability of our proposed evaluation protocol will reach beyond the visual location summarization problem addressed in this chapter and potentially find use in other areas of interest for multimedia community, such as e.g., image set diversification. For example, currently it is very easy to collect a vast amount of web images to be used for algorithm development, but the real problem is the absence of the human-assigned ground-truth needed to determine the relevance of the collected images for the given task. Our proposed evaluation protocol could make it possible to evaluate the specific criteria for general image set representativeness and diversity on a large set of geo-referenced, but not manually labeled data.

3.5.2 New evaluation protocol

Building on the rationale described in the previous section, we propose a simple and intuitive method to evaluate the created visual summaries that makes use of the geotags associated with the images. More specifically, we investigate to which extent the automatically generated visual summary matches the inherent properties of the image collection it represents in terms of the geographical spread of the images in the collection.

We assume that the geographical spread of the images reflects relative popularity and importance of the places within a given area, which needs to be reflected in the visual summary of the area. For example, within the area of 1km in the central districts of Paris or Rome, it is highly unlikely that the images captured by the tourists will have a uniform geographical spread. Instead, the number of photographs taken will be larger at places with multiple “eye catchers” (e.g., a square with multiple monuments and buildings).

To model the geographical spread of the images within a given area, we cluster them based on their geo-coordinates (latitude and longitude), using affinity propagation clustering [22], which automatically determines the number of clusters present. Since the data is low-dimensional, the clustering process is efficient and reliable. We expect our summarization algorithm to output a results list that ”reproduces” the geographical spread of the images in the collection it represents. Namely, to satisfy representativeness and diversity criteria, ideally, all geo-clusters should be represented in the final results list of images, but the number of images in the results set falling into each geo-cluster should also be proportional to the cluster's relative size with regard to the total number of images in the collection. For example, if the geographical clustering detects three clusters, having 60%, 30% and 10% images, respectively, then our list of representative and diverse images, consisting of i.e., $N_R = 10$ images, should have 6 images from the first, 3 from second and 1 from the third cluster.

Having this in mind, we propose to evaluate the summarization results using the multinomial distribution, a generalization of the binomial distribution. First, we define the zero hypothesis using the relative sizes of k detected geo-clusters and consider them as clusters' prior probabilities $p_c = (p_{c_1}, p_{c_2}, \ldots, p_{c_k})$, where $\sum_{j=1}^{k} p_{c_j} = 1$. Let the number of images in the final results set drawn from each cluster $C_j, j = 1 \ldots k$ be $x = (x_1, \ldots, x_k), \sum_{j=1}^{k} x_j = N_R$. Then we evaluate the selected set of representative and diverse images using the outcome of the multinomial distribution probability mass function $f(x|N_R, p_c)$:

$$\begin{aligned} f(x|N_R, p_c) &= \binom{N_R}{x_1, \ldots, x_k} p_{c_1}^{x_1} p_{c_2}^{x_2} \cdots p_{c_k}^{x_k} \\ &= \frac{N_R}{x_1! \cdots x_k!} p_{c_1}^{x_1} p_{c_2}^{x_2} \cdots p_{c_k}^{x_k} \end{aligned} \tag{3.8}$$

The multinomial distribution PMF (cf. (3.8)) has the maximum when the relative number of geo-clusters' observations in the resulting image set corresponds to the clusters' prior probabilities (relative size of detected geo-clusters) - $\forall j \in [1, k] : \frac{x_j}{N_R} = p_{c_j}$. Also, the function f has a slight bias towards dominant (i.e., larger) clusters so that a complete absence of the most dominant geo-cluster in the result set yields relatively higher decrease in $f(x|N_R, p_c)$ score, than the absence of an outlier or a minor cluster.

For the purpose of evaluating our approach, multinomial distribution PMF (MNPMF) is to be preferred since it produces a clearly explainable probability. The evaluation protocol could be also formulated in the way such that a widely deployed alternative, such as the Kullback Leibler (KL) divergence, can be used to evaluate the results. The KL divergence, however, produces unbounded scores which cannot be meaningfully combined by averaging. This is essential in our case, since we need to aggregate performance achieved for multiple locations that have potentially different underlying geographic distribution of images.

3.6 Experimental setup

3.6.1 Image collection

To test our approach we first selected 500 geo-locations in Paris, France, as the outputs of a destination recommender system [12]. For each of the locations we downloaded up to 100 CC-licensed Flickr images taken within the radius of 1km from each location together with the accompanying metadata, i.e., image title, description, tags, geotags, information on uploaders and commentators. Some of the selected geographic areas overlap to a certain degree and, therefore, a single image may appear in more than one collection subset. In order to filter out bad-quality images, we downloaded the images according to their Flickr popularity score. While we reason in Section 3.3 that popularity is not that suitable as a criterion for summary generation, using a reasonable popularity level as a filtering criterion proved to be useful in selecting good image candidates for summary generation. Based on the number of downloaded images per location, we kept only those locations for which 100 Flickr images could be downloaded. We selected the number 100 as a tradeoff between the number of images per location and the number of available locations to work with. Furthermore, 100 images per location proved to provide an appropriate source set for building visual summaries in our case and also to provide a good balance between mainstream and non-mainstream images. Each image was accompanied with the highest-accuracy geotag available in Flickr (accuracy level 16) to make sure that the images with the same or similar geotag were actually captured at the nearby locations. To model implicit user relations, we used images from all available locations.

The requirement for 100 relevant images per location led to 207 locations that we used for our experiments. The selected 207 areas of Paris are often very different, both from the perspective of semantic density of the region and from the perspective of visual homogeneity, which challenges the robustness of our approach. In other words, in case of areas with e.g., a single dominant landmark, user-generated images are expected to be more homogeneous than in the case of areas with multiple dominant landmarks or the areas with no recognizable landmark at all. The variation is independent of the exact radius of the image sampling region and for this reason we expect good generalization of our approach should another radius be used in an operational setting. For each of the selected 207 locations we constructed a graph and selected the representative and diverse images as described in the previous section.

3.6.2 Baseline approaches

To evaluate the effectiveness of our approach with respect to the related work, we compare our algorithm with six baseline methods:

Random: Random selection of N_R images per location.

View Count: N_R images with the highest "view count" are selected (number of times images are viewed).

Number of Comments: We conjecture that the number of comments is a useful indicator of the "image appeal" in Flickr. Here we select N_R images with the largest number of comments to summarize given location.

Visual Clustering: K-means clustering is applied to visual feature representations to cluster images into N_R clusters. The image closest to the cluster centroid is selected to represent each cluster. We adopted k-means as the clustering mechanism to be consistent with similar work on cluster-based approaches carried out in the past, such as [41]. We have also experimented with other clustering approaches, like hierarchical clustering and affinity propagation clustering [22]. Although the number of images per location is relatively small (100), due to a high visual diversity of user generated content and the high dimensionality of the feature space, these more recently proposed clustering algorithms do not provide appreciable improvement over k-means clustering.

MA Clustering: As the basis for clustering we use multimodal similarities (affinities) between image nodes computed as described in Section 3.4.2 and stored in matrix $\mathbf{S} = [s_{lj}]_{N \times N}$. Images are clustered using a well-known affinity propagation clustering approach [22]. In the related work [69], [9], affinity propagation clustering was proven effective in tasks similar to ours and we use it here also because it does not require feature vectors as input, but rather similarities between data points. After the N_C image clusters are created, we sort them in the descending order in terms of the number of elements. In case of this particular collection, the number of detected clusters N_C varies between 11 and 28 depending on the geographic area. If N_C is larger than or equal to the size N_R of the visual summary to be created, we create the visual summary by simply sampling the exemplars [22] of N_R top-ranked clusters. Otherwise, we first select the exemplars of the N_C detected clusters and further select $N_R - N_C$ remaining images in an iterative fashion. We start with the top ranked cluster and select its centroid, or in other words, amongst those images that have not already been included in the visual summary we select the one with the highest average multimodal similarity with the other images in the cluster. We proceed by sampling the next ranked cluster in the same manner until N_R images are selected. Besides serving as the general visual summarization baseline, this approach is also intended to confirm the effectiveness of our proposed diversification strategy. Namely, as discussed in Section 3.2, image clustering and selection of cluster representative for the final results set is a common diversification strategy [41, 106].

Cluster Ensemble: We conjecture that the use of multiple modalities might improve clustering performance significantly. First, similar to Visual Clustering described above, we make use of K-means clustering to group images together into exactly N_R clusters. Images are clustered independently based on their visual and text features. To combine individual clustering results into a single consolidated clustering, we apply the cluster ensemble framework [100]. Strehl and Ghosh [100] demonstrate that in various scenarios cluster ensemble (in literature also known as consensus clustering) yields results that are at least as good as the results of any individual clustering being combined. After the N_R clusters have been created, we create the final visual summary by collecting images corresponding to clusters' visual centroids.

3.7 Experimental results

We perform a set of experiments to answer the following research questions:

1. Is our approach, denoted by RWR-RD capable of selecting a good set of representative and diverse images to create an informative visual summary of a particular geographic area?

Table 3.1: Performance of the proposed RWR-RD approach and the 6 baseline approaches with regard to representativeness and diversity of the selected image set; the performance is reported in terms of MNPMF averaged over all 207 locations

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
Random	0.0213	0.0116	0.0066	0.0045
View Count	0.0171	0.0084	0.0055	0.0039
Nr. of Comments	0.0030	0.0006	0.0001	0.0002
Visual Clustering	0.0154	0.0039	0.0014	0.0017
Cluster Ensemble	0.0204	0.0102	0.0052	0.0047
MA Clustering	0.0263	0.0125	0.0055	0.0035
RWR-RD	**0.0320**	**0.0161**	**0.0092**	**0.0062**

2. Is the performance stable across locations?

3. Does combining modalities, namely visual features, text and user relations, improve the performance over using modalities individually? Can knowledge of the contribution of individual modalities be exploited to further improve the proposed approach?

4. Is our approach capable of addressing the long tail problem and selecting non-mainstream (off-the-beaten track) images?

The following subsections address each of these questions in turn.

3.7.1 General performance evaluation

In the first experiment we compare our approach that we denote as RWR-RD with the 6 baseline approaches described in the previous section: random, view count, number of comments, visual clustering, MA clustering and cluster ensemble. The performance is evaluated using multinomial distribution PMF (MNPMF) described by (3.8). In Table 3.1 we report the results averaged over all 207 location in Paris for which we generated visual summaries. No additional weighting is applied to attribute and similarity edges in the graph (i.e., modality-dependent weights are set to $\beta_f = \beta_t = \beta_u = 1$).

As shown in Table 3.1, our proposed RWR-RD method clearly outperforms all baselines in terms of representativeness and diversity for various sizes of the set of selected images, $N_R = 5, 10, 15, 20$. Further, view count and number of comments approaches, commonly used in commercial systems, were found to perform poorly. When choosing among the light-weight selection approaches, including view count, number of comments and random sampling, for better geographic representativeness and diversity we would suggest pulling images randomly (possibly in combination with view count). This option is also intuitive, because in the case of a large result set, random selection would yield frequency of geo-clusters proportional to their relative size.

As expected, multimodal clustering approaches, MA clustering and cluster ensemble outperform unimodal visual clustering. Further, MA clustering emerges as the overall second best performer, which again confirms effectiveness of the proposed graph structure and the approach for computing the multimodal image similarities. Since both MA clustering and RWR-RD approaches utilize the same multimodal image affinities computed from the graph, a significantly higher performance of RWR-RD approach demonstrates the added benefit of our proposed iterative procedure for selection of representative images and maximization of set diversity.

The results in Table 3.1 also reveal that the performance of all tested approaches drops with the increasing size of selected image set N_R. However, in our use scenario we intend to use the visual summaries composed of only a limited number of images (e.g., 5-10), which makes results in the left part of Table 3.1 more relevant in the context of this chapter and, consequently, our proposed RWR-RD method more suitable for the task.

The results are in line with the preliminary user study in which our algorithm was evaluated as a part of a larger destination recommender system [42]. Majority of participants appreciated the level of diversity in the automatically created visual summaries. They received a good overview of the presented geographic area and most of them also had the impression that the images were selected carefully.

3.7.2 Performance stability across locations

Although measuring the average performance over multiple queries is common practice [64, 72], it may also be misleading, especially in the cases when exceptionally large improvement is obtained for a limited number of queries, while for many other queries the performance deteriorates.

Therefore, in this experiment, we choose to compare the performance of our proposed RWR-RD approach to the baselines for each individual location. Table 3.2 shows percentages of locations in which a particular

Table 3.2: Performance of our RWR-RD approach and the 6 baseline approaches with regard to representativeness and diversity of the selected image set; percentage of locations for which a particular approach is the best performer with respect to the MNPMF score is reported

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
Random	14.0	10.6	11.6	15.0
View Count	9.2	5.3	3.9	2.8
Nr. of Comments	4.8	2.4	1.9	1.4
Visual Clustering	11.1	10.6	11.6	14.5
Cluster Ensemble	9.7	16.9	12.1	15.0
MA Clustering	**28.0**	17.0	19.8	18.4
RWR-RD	23.2	**37.2**	**39.1**	**32.9**

method yields the best results. An additional advantage of this way of evaluating the algorithm instead of using the values of MNPMF explicitly, is that it corresponds to a typical way users are asked to compare the quality of image sets, namely by telling which one of several options better suits their needs. As shown in Table 3.2, our RWR-RD method clearly outperforms the baseline approaches in terms of representativeness and diversity for most N_R settings.

3.7.3 Modality analysis

In Table 3.3 we compare the performance of our multi-modal RWR-RD approach (all modalities used) with the same approach exploiting only a single modality at a time, namely visual features (RWR-RD-F), text (RWR-RD-T) or interactions of the users and their social network with the images (RWR-RD-U). As discussed in Section 3.4.1, the three unimodal alternatives were realized by replacing the graph G by the subgraphs G'_m, whose node sets are $V'_m, m \in \{f, t, u\}$.

Table 3.3 reveals that our RWR-RD method benefits from the use of multiple modalities, since in most cases not a single modality achieves the performance of the multi-modal case. The algorithmic simplification that exploits only the interactions between users and images and users and their social network (RWR-RD-U) shows surprisingly good results in producing representative and diverse visual summaries. A logical explanation for this effect might be that different users capturing images in a particular

Table 3.3: Analysis of the contribution of each modality to the performance of our RWR-RD approach; the performance is reported in terms of MNPMF averaged over all 207 locations

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
RWR-RD-F	0.0279	0.0125	0.0083	0.0055
RWR-RD-T	0.0091	0.0051	0.0038	0.0030
RWR-RD-U	**0.0329**	0.0147	0.0088	0.0058
RWR-RD	0.0320	**0.0161**	**0.0092**	**0.0062**

Table 3.4: Performance of our summarization approach with and without modality-dependent weights; the performance is reported in terms of MNPMF averaged over all 207 locations

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
RWR-RD-W	**0.0339**	**0.0165**	**0.0101**	**0.0067**
RWR-RD	0.0320	0.0161	0.0092	0.0062

area often have different interests and thus generate, to a certain degree, unique photo-streams covering various aspects of the location. Also, it is reasonable to assume that the people commenting on each others' images might have similar tastes and our approach is designed to exploit these relations. The relatively good performance and low computational complexity of RWR-RD-U approach suggest that it could be efficiently applied in the cases where the computational load is a critical parameter. The lower performance of the text-based RWR-RD-T algorithm variant might be an artifact of users' tagging behavior, such as, e.g., adding personal tags or the tendency of different users to assign very similar sets of tags to the images captured within a particular area (i.e., the tag set is often non-discriminative).

Based on the outcomes of modality analysis, we compute modality-dependent weights as described in Section 3.4.1 and apply them to modify the edges in graph G. We use the performance of unimodal approaches (i.e., π_f, π_t and π_u) for $N_R = 15$ to compute these weights. In Table 3.4 we compare the performance of our standard RWR-RD method with the one utilizing modality-dependent weights RWR-RD-W.

The results in Table 3.4 show that the use of modality-dependent weights brings consistent additional improvement of the performance. Al-

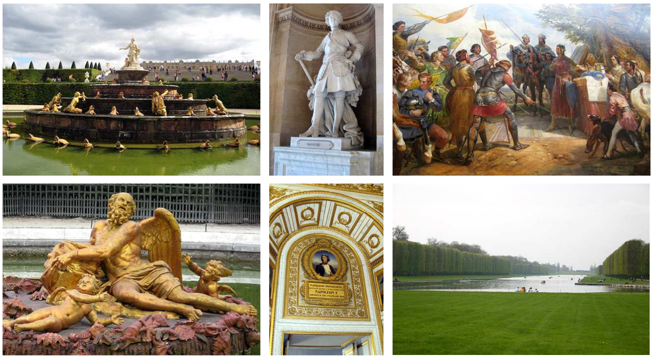


Figure 3.6: Example visual summary for a location in the "Château de Versailles" area automatically generated by our RWR-RD-W approach; the images show in the clockwise order Latona Fountain with the palace in the background, statue of Louis XV, The Battle of Bouvines painted by Horace Vernet, The Grand Canal, Napoleon's portrait and the Saturn Fountain. All images are downloaded from Flickr under CC license.

though this improvement is not always substantial, the results support the insight related to the importance of modality-dependent edge weighting as discussed in Section 3.4.1.

Fig. 3.6 shows the example visual summary generated by our RWR-RD-W approach for a location in the Château de Versailles (Versailles Palace) area. The images illustrate various aspects of the area that could be interesting for a potential visitor and range from fountains, statues and historic paintings to popular resting spots such as e.g., The Grand Canal. An ideal summary would probably also include the image of the Palace building, which is here featured in the background of the first image, but surprisingly, only a very few (compared to the other semantic categories, e.g., fountains, paintings or statues) out of 100 images in the initial set have the Palace façade as the dominant object. Therefore, they have not been found representative enough to appear in this compact summary.

3.7.4 Selecting non-mainstream images

In this experiment we evaluate whether our approach is capable of selecting non-mainstream images as well. As the non-mainstream images we consider those that belong to the "long tail" of less popular, but potentially interesting part of image collection taken in the considered area. Most commercial systems still fail to present these images to the users and for this reason we expect them to have in average smaller number of comments.

Table 3.5: Performance of our RWR-RD-W approach and the 4 baseline approaches with regard to their ability to select the non-mainstream images as well; percentage of locations for which a particular approach is the best performer is reported

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
Random	24.2	18.7	19.3	19.8
Visual Clustering	11.6	9.7	7.8	5.8
Cluster Ensemble	16.4	18.4	16.4	14.5
MA Clustering	11.1	6.8	5.3	5.3
RWR-RD-W	**36.7**	**46.4**	**51.2**	**54.6**

In order to investigate the effectiveness of our approach with this respect, we take the average number of comments per image in the generated summary as the evaluation criterion and compare our method with random, visual clustering, cluster ensemble and MA clustering baselines. Table 3.5 shows the percentage of locations for which a particular approach selects the image set with lower average number of comments per image than the other four approaches. We do not use view count baseline because of a high correlation between nr. of comments and view count.

Our RWR-RD-W approach clearly selects largest number of non-mainstream images in more locations than the baselines.

3.8 Discussion

Experimental results presented and analyzed in the previous section indicate the effectiveness of our approach in generating a compact set of representative and diverse images of an area. It outperforms the baseline approaches both in terms of average performance over 207 selected locations as well as percentage of locations for which it performs better then the baselines. Additionally, we demonstrate the benefits of using multiple modalities and also propose favorable settings (e.g., RWR-RD-U) for the visual summarization task in cases where computational complexity is a critical parameter. The experiments reveal that our approach is capable of selecting not only mainstream images, but less popular images as well, which would otherwise stay lost in the long tail of items which are potentially valuable, but inaccessible in practice.

Although our approach emerges as the best performer for the task addressed in this chapter, for some locations it performs worse than some of the baselines. In our future work we will further examine the factors that negatively affect the algorithm's performance. Our initial research on automatic computation of modality-dependent and location specific weights for improved summarization performance shows promising results. Our future work will include analysis of pair-wise unimodal image similarities, aiming at the development of potentially better and computationally less intensive modality-dependent weighting approaches. We will work on developing more sophisticated methods for the optimization of results with respect to representativeness and diversity criteria, especially those reflecting users' preferences, an aspect which will require careful planning of large scale user tests.

To complement the results presented in the experimental section of this chapter, we will investigate the topical coverage of the resulting image set produced by our proposed approach and the baselines. Exploratory experiments have shown that compared to the baselines our RWR-RD approach selects images whose tags are more representative of a given area (e.g., tags associated with many images captured within that area). We leave a deeper analysis of this observation for the future work. Currently we use geotags only for the evaluation of our results. In the future, we will also analyze a scenario in which very precise geotags are available for all images (or at least significant number of images) captured within a particular area and use them to improve results of our visual summarization approach.

Finally, we will design and carry out large-scale user tests in the form of a crowdsourcing task on a platform such as Amazon's Mechanical Turk. The tests will aim at investigating the effectiveness of our algorithm in fulfilling user needs with respect to the user-perceived aesthetic value of the generated visual summary, e.g., in terms of appeal or sentiment. In addition, the tests will be designed to help us gain information from the users that could steer our search for the criteria for improving the aesthetic quality of the visual summaries and for mapping these crtieria into concrete feature-based image representations and analysis approaches.

Chapter 4

User-informed visual summarization

In the previous chapter, our main goal was to create visual summaries depicting all relevant aspects of a geographic area, making sure that the selected images are the most representative of the aspects they are illustrating. Here we make use of crowdsourcing to get an insight into how humans perform visual summarization. Based on the outcomes of the user study, we propose a novel algorithm which utilizes a heterogeneous feature representation to learn to rank images according to their suitability for visual summarization. Finally, we propose an evaluation protocol tailored to optimally exploit the properties of human-created visual summaries and inspired by the metrics used in evaluation of textual summaries.

This chapter is accepted for publication with mandatory minor revisions as: Stevan Rudinac, Martha Larson, Alan Hanjalic. Learning crowdsourced user preferences for visual summarization of image collections. *IEEE Transactions on Multimedia.*

4.1 Introduction

Rapid growth of the amount of digital multimedia data available in personal and professional collections as well as the content sharing and social networking websites, has created the need for powerful tools enabling analysis, representation, abstraction and summarization of data for more efficient and effective browsing and retrieval. Summarization techniques, in particular, aim at providing a compact representation of a single multimedia data document or data collection. Depending on the type of data and the application domain, summaries may consist of text, images, video segments or a combination of these.

In this chapter we focus on visual summaries. Visual summaries serve to abstract a video [65, 94], set of videos [48] or an image collection [9, 41, 81] and usually consist of video segments or images (e.g., photos or video keyframes).

Although humans in general intuitively understand the concept of a (visual) summary, giving a single and universal definition of the summary appears to be difficult [77]. While intuitively the structure and content of a summary should depend on the purpose it should fulfill [25], the final assessment of its quality can only be made based on its compatibility with the expectations of the human users. Therefore, given a particular application and use case, the specific criteria reflecting the user's perception of the summarization quality should be identified and used to steer the summarization algorithm. In other words, a summarization algorithm should be *user informed* in order to be successful.

Existing methods for visual summarization have typically been guided by studies (e.g., [10]) of users' preferences in terms of a tradeoff between the relevance and representativeness of the information included in the summary and the ability of the summarization algorithm to diversify the included visual content [9, 41, 69, 81]. The notions of relevance, representativeness and diversity, as well as the interplay among the three are, however, too general to be modeled successfully in a given summarization scenario, and especially across scenarios. Furthermore, although the quality of visual summaries generated using the existing approaches is sometimes judged by human evaluators (e.g., [41]), explicit information on how humans create visual summaries has hardly been inferred or taken into account while developing summarization algorithms. Therefore, the insights obtained so far can be considered insufficient to serve as guidelines for developing a solid visual summarization approach.

In this chapter we demonstrate how user-informed visual summarization algorithms can be facilitated by relying on *crowdsourcing*. We first run a large-scale crowdsourcing experiment to obtain insight into how users perform visual summarization. Then we use this insight to decide on the appropriate features, based on which images in the collection can be ranked. The ranking reflects the suitability of an image as a candidate for inclusion in the summary, that is, how likely an image would be selected for the summary by the users.

We take the problem of visual summarization of geographic areas as the sample use case in this chapter to demonstrate the benefits of the proposed user-informed image selection concept. We foresee, however, that the material presented here will be of use in a wide range of summarization problems. The chapter makes the following main contributions, whose implications transcend our specific choice of use case:

- We show how to deploy crowdsourcing to acquire implicit and explicit criteria humans find important when performing visual summarization.

- We propose a novel approach for embedding the derived criteria into descriptive features and learning to distinguish between images based on the likelihood of their appearance in the human-created visual summaries.

- In order to match the criteria inferred from human-created summaries, we expand the scope of features used to represent the image collection beyond those that are typically deployed for visual summarization. This expansion encompasses, in particular, features related to the context, aesthetic appeal, sentiment and popularity of an image.

- We provide new insights regarding the applicability of some standard image aesthetic appeal features in a general summarization scenario.

- We demonstrate that the existence of multiple "optimal" visual summaries leads to a low inter-user agreement that makes image set evaluation difficult. We therefore propose an automatic evaluation protocol based on the pyramid approach and motivated by the experience from the text domain that has been documented in the literature by the text summarization and machine translation communities.

In Section 4.2 we provide an overview of the proposed image selection approach and explain in more detail the rationale behind it. In Section 4.3 we report on related work. In Section 4.4 our crowdsourcing experiment is described and then in Section 4.5 we present the features used to represent images. Our approach to user-informed image selection is introduced in Section 4.6. Section 4.7 details the pyramid approach to summary/image set evaluation, while in sections 4.8 and 4.9 we present the experimental results. Finally, Section 4.10 concludes the chapter.

4.2 Approach overview and rationale

Our approach to user-informed image selection for the purpose of summarizing an image collection is illustrated in Fig. 4.1. To allow us to develop a deeper understanding of how people create summaries of image collections, we first run a crowdsourcing experiment on the Amazon Mechanical Turk[7] platform and collect a large number of manually created visual summaries. The participants of the study were also asked to indicate the reasons for selecting a particular image for the summary, which helped us acquire insight into the general criteria that should be satisfied by an automatic summarization algorithm.

In the next step, we map these criteria on a number of features used to represent the images in the collection, both in terms of their individual properties and in the context of other images in the collection. Feature selection is steered by two main observations derived from the crowdsourcing experiment. First, we observed that the number of semantically related images in the original collection plays an important role when selecting an image for the summary (e.g., related to the paradigms of diversity and representativeness as introduced in the previous work [41, 81]). We consider images to be semantically related if they are captured at nearby locations (e.g., having the same or similar geo-coordinates) and are also visually similar to each other (e.g., depict the same scenes, objects or events). Images captured at the same geo-location, but with different depicted content are considered semantically different. Based on this understanding of semantic similarity, we consider geo-coordinates and standard images features, which reflect the saliency of the depicted visual content (object, scene) as the input for geo-visual clustering that reveals semantic links among the images in the collection.

[7] https://www.mturk.com/mturk/welcome

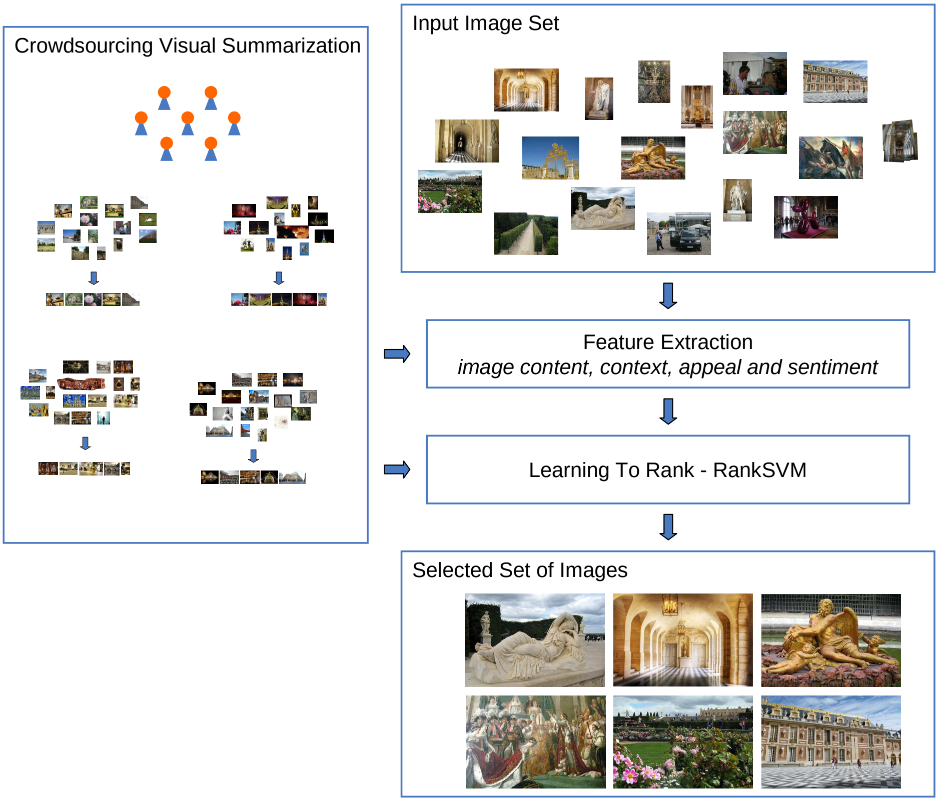

Crowdsourcing Visual Summarization
Input Image Set
Feature Extraction
image content, context, appeal and sentiment
Learning To Rank - RankSVM
Selected Set of Images

Figure 4.1: Illustration of the proposed user-informed approach to image selection for creating visual summaries. All images are downloaded from Flickr under CC license.

We observed, however, that some other more subtle criteria also played an important role when the human summarizers were deciding on which images to select for the summaries. While typically a low inter-user agreement is expected regarding the inclusion of a specific image in a summary (probability is inversely proportional to the number of equally qualifying candidate images), it was striking to see that some images were selected by many users, far more often than other images. Based on the comments the users provided with their summaries, we concluded that an explanation of the criteria for image selection in these cases could be linked to the notions of *image aesthetic appeal* [17, 54, 62], *affect* and *sentiment* [93, 111] that have been investigated in various research contexts, such as e.g., image processing and computer vision, affective computing, natural language processing and social network analytics.

Therefore, similar to e.g., [62] we extract several image aesthetic appeal features (e.g., image colorfulness, aspect ratio) and consider image popularity indicators as well (i.e., view count and number of comments). For consistency reasons, we adopt notation from related work, where image aesthetic appeal features are considered to be those that influence aesthetic rating of an image [17, 54, 62]. Regarding the sentiment, similar to [93] we conjecture that the useful information might be derived from the comments posted on images, which often have an affective dimension. For the reasons of consistency with the related work, we refer to this particular step as the image sentiment analysis. Our sentiment analysis approach is based on publicly available Whissell's Dictionary of Affect in Language [110], attempting to quantify emotions in natural language. Finally, we investigate whether the targeted levels of appeal and sentiment can also be detected indirectly using various popularity indicators that can be derived from popular online image sharing sites.

The selected features serve as input into our proposed image selection approach. This approach aims at learning inherent properties that make images more or less likely to be selected for the summary by humans. We start from the reference summaries obtained through crowdsourcing and train a RankSVM [11] for each collection subset, providing frequently selected images as the positive and least frequently selected images as the negative examples. The final image ranking, which could be used as input when producing a visual summary, is generated by rank aggregation as explained in detail in Section 4.6.

4.3 Related work

In this section we discuss previous work related to the problems and technologies addressed in the chapter.

4.3.1 Visual summarization

Generally, visual summarization aims at building a compact representation of a single video, set of videos or an image collection. Informedia [109] was probably one of the earliest projects addressing video summarization. More recently, TRECVID benchmark series run the BBC rushes summarization evaluation pilot (e.g., [65]), where the benchmark participants were provided 40 BBC rushes video files for each of which they were expected to generate visual summaries with up to 2% of the duration of the original file.

With the growing popularity of social media, a number of approaches for generating summaries of collections of community-contributed images have been proposed. Kennedy and Naaman [41] propose a multimodal approach to providing representative and diverse views of landmarks using Flickr images. In [69] travelogues and Flickr images are used for creating the summaries of touristic cities. Popescu et al. [75] make use of Flickr images and associated metadata for discovery and recommendation of tourist trips. Cao et al. [9] first cluster Flickr images using associated geo-coordinates and then represent each geo-cluster by the most representative images and the most frequent tags. In our previous work [81] we presented a multimodal approach to visual summarization of geographic areas using community contributed images. The approach makes use of visual content of the images, associated annotations (i.e., title, description and tags) as well as the information about users and their social network to select representative, but diverse images of a geographic area within a predefined radius from a selected location.

Visual summarization of data recorded by the wearable capturing devices is another example of application domain rapidly gaining popularity in the research community. For example, given a video recording of a wearable camera, Lee at al. [46] propose an approach, which jointly utilizes saliency detection and temporal event analysis for automatically generating visual summaries depicting the most important people and objects appearing in the video.

4.3.2 Summary evaluation

Automatic summary evaluation has been a topic of intensive research in the (text) information retrieval community [25, 77] and although many different metrics have been proposed over the years, the evaluation problem still poses significant challenges. Since 2001, the Document Understanding Conference (DUC) series [16] and the successor series, Text Analysis Conference (TAC) have been the epicenter of research in the field of automatic summarization and summary evaluation [66]. The majority of the proposed metrics for summary evaluation have relied on the assumption that a good summary should be as similar as possible to one, or preferably more, human-created reference summaries. In [71], BLEU, an algorithm for automatic evaluation of machine translation was proposed. The main idea behind BLEU is to compare candidate translation with several reference translations (e.g., translations made by humans) using n-gram co-occurrence statistics. ROUGE [49] is another well-known example of the metric for evaluation of machine translation and automatic summarization, based on a similar idea.

A common problem with the automatic summary evaluation metrics such as e.g., ROUGE is a low agreement between human-created reference summaries. Therefore, based on the assumption that some summarization content units (SCUs) are more important and therefore should be given a higher weight when scoring summaries, a pyramid evaluation approach was proposed [60] and later adopted by TAC as the official summary evaluation metric [66]. Although it shows a high correlation with the human judgment about the quality of an automatically generated summary, the pyramid approach has the drawback that the SCUs need to be manually annotated.

Compared to the field of document summarization, the multimedia community has made relatively few attempts to systematically evaluate visual summaries. The participating video summarization systems in TRECVID BBC rushes benchmark [65] were evaluated using common metrics. However, the automatic evaluation was not the focus of the initiative and the summaries were judged on several parameters by the human evaluators.

Inspired by the well known BLEU [71] and ROUGE [49] metrics, Li and Merialdo [48] proposed VERT, an algorithm targeting automatic evaluation of video summaries. While BLEU and ROUGE compare candidate summary with several human-created reference summaries in terms of e.g., n-gram co-occurrence statistics, as a unit for comparison VERT analogously uses the "group of n keyframes" as an alternative. However, as will be illustrated in Section 4.7, a very low overlap between human-created reference summaries deems the evaluation metrics such as BLEU, ROUGE and VERT inapplicable to the task addressed in this chapter.

4.3.3 Image aesthetic appeal and sentiment analysis

Estimating image aesthetic appeal as well as the sentiment that images evoke is a complex problem that has been a subject of intensive research. Approaches to image aesthetic appeal estimation aim at measuring the image properties that make it appealing to the user. In [91] a user study was conducted to identify those properties, which led to several categories of features related to e.g., people, composition/subject, quality (blur, contrast etc.) and redundancy. Example image properties found by the similar studies to be correlated with the aesthetic appeal include image colorfulness, sharpness, rule of thirds, size, aspect ratio and face appeal features amongst many other [17, 26, 54, 62, 112].

As a result of the increased popularity of social media in the recent years, the analysis of sentiment evoked by multimedia content is becoming increasingly more sophisticated and easier to carry out. For example, from the comments posted in relation to a YouTube video or a Flickr image, it is often possible to understand whether users perceive the multimedia item as e.g., pleasing, happy or sad. Recently, the publicly available lexical resources such as e.g., Whissell's Dictionary of Affect in Language (DAL) [110] and SentiWordNet [21] have been proven effective in sentiment analysis of digital content. In the process of the creation of the DAL, a large number of words were annotated with regard to their *pleasantness (valence)*, *activation* and *imagery*. Similarly, in SentiWordNet, each synset of WordNet lexical database [53] is accompanied by *positivity*, *negativity* and *objectivity* sentiment scores. For example, in [37] DAL was successfully utilized for detection of narrative peaks in documentary videos, while in [92] SentiWordNet was deployed for predicting the rating of YouTube comments. In another recent study, Siersdorfer et al. [93] make use of SentiWordNet to analyze user-generated comments associated with the Flickr images and quantify their sentiment.

4.4 Crowdsourcing for visual summarization

Our automatic image selection approach is informed by the large-scale user tests, which are carried out to investigate the criteria that guide user's selection of images for the visual summary. Below we first describe the image dataset used in the study and then elaborate on the setup and the lessons learned from the crowdsourcing experiment.

4.4.1 Image collection

For the user tests we make use of Flickr image collection described in detail in our recent work [81]. We initially selected 500 geo-locations in Paris, France output by a location recommender system [12] and downloaded at most 100 creative commons (CC) licensed images captured within 1km of each location together with the associated metadata such as e.g., title, keywords, description, comments, geotags (latitude and longitude), information on uploader and commenters. Finally, we kept only 207 locations for which 100 images were available. Downloaded images were selected based on a high Flickr popularity score, which ensures reasonable quality and relevance. The images were not pre-filtered according to the type or topic and thus reflect a wide spectrum of users' interests, such as e.g., landmarks, various types of events in both indoor and outdoor setting as well as the people in their everyday activities. Underlying variations in semantic density and visual homogeneity of 207 selected locations have a similar effect as varying the area size or sampling a varying number of images.

4.4.2 Crowdsourcing experiment

Recently, crowdsourcing platforms such as e.g., Amazon Mechanical Turk (MTurk) and CrowdFlower[8] have emerged as the powerful tools for efficient and relatively inexpensive completion of tasks that require human intelligence. On MTurk, such tasks are called Human Intelligence Tasks (HITs) and can take various forms, such as e.g., translating text from one language to another, rating or tagging images, videos and music. While in the beginning, the majority of MTurk workers were US-based, the recent studies suggest a rapid internationalization of the MTurk labor force [78]. A number of studies have shown that with appropriate design of the HIT, a crowdsourcing platform will yield the same annotations or answers as conventional approaches for collecting judgments from users, e.g., in a laboratory setting [61, 70]. Since the crowdsourcing is a relatively young discipline, to assure a high quality of results and avoid spamming, the HIT design should be approached carefully. Namely, as suggested by [39] the quality of results depends on the factors such as e.g., payment amount per HIT, task complexity and worker qualification/reputation. However, the same study suggests that there is no universal recipe on how to choose those parameters. For example, increasing payment per HIT generally results in a higher quality of results, but it also attracts workers with a more sophisticated spamming methods. Similarly, while increasing the task complexity (effort) might lead to a higher amount of spam, it also yields a higher quality of results after the spam is removed. The study presented in [20] investigates techniques that help detect malicious workers and consequently reduce amount of spam. For example, the study suggests that the malicious users are less inclined to accept tasks involving free text inputs than e.g., those with check boxes.

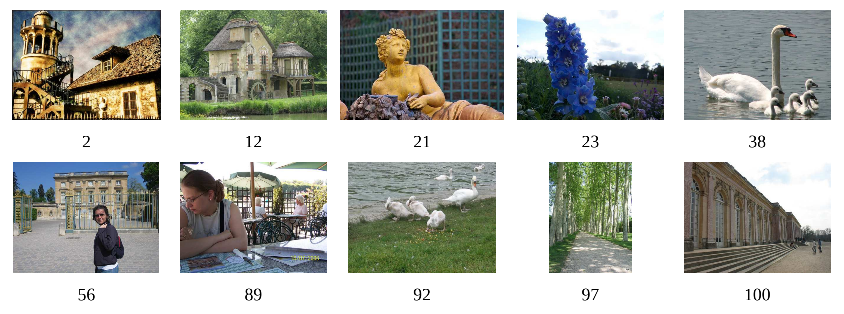
[8] http://crowdflower.com/

Considering these and related recommendations for ensuring a high quality of results, we designed our crowdsourcing task as follows. We recruited 20 different MTurk workers per location for manual creation of reference summaries. As some of them repeated the HIT for the other locations as well, the total number of workers used for the task was 697. The images of a given location were displayed to the worker in 10 rows with 10 images each. To get a better overview of the entire location and fit images to the width of a computer screen, the height of each displayed image was set to 60 pixels. The workers were able to scroll vertically and horizontally and click on the image to see it in full resolution. In the task description, we avoided steering the workers towards any specific criteria for summary creation or to bias them by revealing information about the location. The precise wording of the task was: "*In this task we will show you a set of 100 images and ask you to select 10 of them for a "visual summary". The summary should capture the essence of the larger 100-image set. In other words, by looking at the 10-image visual summary, you should gain the same overall impression as given by the larger 100-image set.*"

After the 10-image summary was created, the worker was asked to sort the selected images in the order of importance and briefly explain reasons for selecting each image using a free text input form. Beside helping us to understand the criteria for summary creation, sorting images in the order of importance and providing reasons for image inclusion in the summary using the free text form served also as another spam control mechanism. Further, the worker was expected to answer several questions about the properties of the original 100-image set, such as e.g., whether it was difficult to create summary of a given image set, whether the presented images in worker's personal opinion show significant or important things and whether they are diverse. The answers were provided via a 4-point Likert scale. Finally, the free form text input was left for the feedback on task complexity, user friendliness, question ambiguity etc. An example visual summary made by a worker is shown in Fig. 4.2.

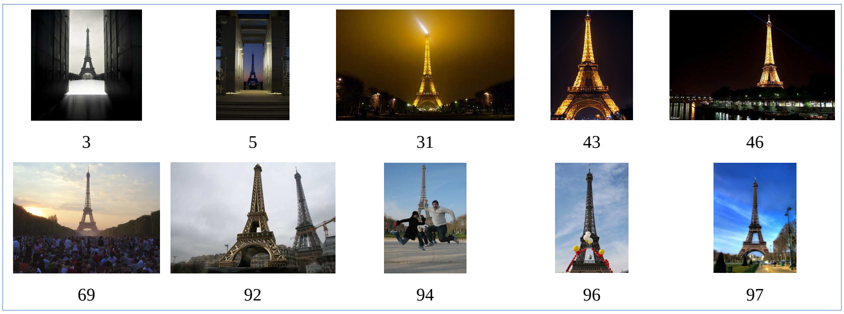
4.4.3 How do users approach visual summarization

We first perform qualitative analysis of the manually generated visual summaries as well as the criteria for image selection reported by the MTurk workers. The analysis reveals that most of them select images that are semantically similar to many other images in the collection, making sure



38	There are several pictures of the swans, and this shows an adult with many children in the water
2	This shows a great work of public architecture
21	This shows an important sculpture
23	This shows the native flora of the land
12	This shows a great architectural work which appears to be a residential house
56	This is one of the tourists next to a landmark
97	This is a great picture, photographically speaking. Shows nature from a great angle
100	Another great photographic picture, which also includes an architectural landmark
89	Shows one of the tourists in a public restaurant or cafe
92	Shows the swans again, only this time some are on land

Figure 4.2: An example visual summary manually generated by an MTurk worker. The images are further sorted in order of importance and the reasons for their inclusion in the summary are indicated.



97	Classic image of the tower, day
92	Closer, different angle, day
3	Framed, day shot of the full tower
5	Framed, evening shot of the full tower
94	Enjoying the sights, day shot
69	Nice shot with the horizon
31	Evening shot fully lit
43	Closer evening shot
46	Nice full shot with the lights in the evening
96	Picture of a celebration, decorations

Figure 4.3: An example of behavior exhibited by a smaller number of workers to represent a particular collection by the images of its most dominant/representative landmark or event.

at the same time that as many semantically different images as possible are included in the summary. In this respect, this observation is in line with the previous user studies such as e.g., [41] and suggests that a trade-off between representativeness and diversity was targeted by the workers. However, we avoid making such explicit hypotheses in this chapter as the analysis also revealed that humans often have distinct and individual perspectives on representativeness and diversity. Imposing the general expectations on a summary and using them to steer the design of a summarization algorithm would therefore be rather artificial and distract the summarization approach from reaching its goal. As an example we compare the summaries in Fig. 4.2 and Fig. 4.3 that have both been generated from the sets of highly diverse images showing various objects and events. Since the worker in Fig. 4.3 decided to include images of the Eiffel Tower only, considering exclusively representativeness and diversity as defined in the previous work would lead to intuitive conclusion that the worker does not consider diversity as an important criterion and that this summary is qualitatively worse than the one in Fig. 4.2. However, the worker in Fig. 4.3 does consider diversity, but at another semantic level (e.g., different views are selected, the images are captured during daytime and nighttime etc.). Such behavior is more frequently observed in the case of image collections including images of well-known objects or events (cf. Fig. 4.3).

Furthermore, we observed that the semantically similar images (e.g., showing the same object or event) were not necessarily considered by the workers as equally suitable for inclusion in the summary. For example, in a particular location for which a summary is shown in Fig. 4.2, 7 out of 100 images are depicting swans. As shown in Fig. 4.4, one of those images was included in the visual summary by 7 (out of 20) workers, which indicates an unusually high consensus (inter-annotator agreement). As already indicated in Section 4.2, we relate this to the notions of image aesthetic appeal and sentiment, which we assume to have influenced the workers during the summary creation.

4.5 Feature extraction

Based on the insights derived from the study in sections 4.3 and 4.4, we propose an approach for automatic user-informed selection of images serving as input for visual summary. Here we first describe the categories of features we extract from the images in the collection. Then, in Section 4.6,

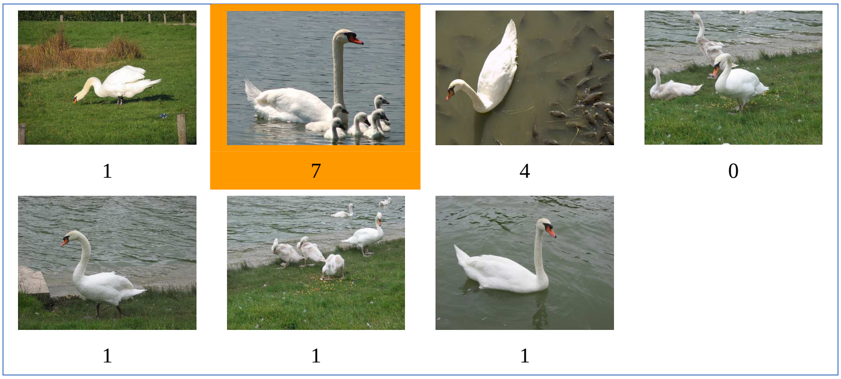


Figure 4.4: An example showing several semantically related images captured in the area around a particular location. The numbers below each image indicate how many out of 20 workers selected that particular image for the visual summary.

we elaborate on the algorithm that deploys these features to learn to rank the images based on their suitability for the visual summary.

As mentioned in the introduction, one of the particular novelties of our approach is that we do not describe each image based on its properties only, but also in the context of the other semantically related images from e.g., the same geo-visual cluster. More particularly, we represent each image i with a feature vector $\mathbf{x}_i$ based on its "importance", popularity, aesthetic appeal and sentiment evoked in the users, but also with the mean and standard deviation of those features computed for the images within the same geo-visual cluster.

4.5.1 Geo-visual clustering

For each of 207 geographic areas (c.f. Section 4.4.1), similar to [9], we first cluster images using their geo-coordinates. To cluster images into a certain number of geo-clusters, we make use of affinity propagation clustering [22], which was proven effective for the similar tasks in our previous work [81] as well as in [9] and [69]. Another property that makes the affinity propagation clustering preferable to some alternatives is its effectiveness in automatically determining the number of clusters.

The inputs into affinity propagation clustering are the similarities between images computed as

$$\mathbf{S}_g(i,j) = \mathrm{sim}(\mathbf{g}_i, \mathbf{g}_j) = e^{-\delta(lat_i, lon_i, lat_j, lon_j)} \tag{4.1}$$

where $\delta(lat_i, lon_i, lat_j, lon_j)$ is the great circle distance between geo-locations $\mathbf{g}_i = (lat_i, lon_i)$ and $\mathbf{g}_j = (lat_j, lon_j)$ associated with the images i and j.

After the geo-clusters are created, we produce the final geo-visual clusters by clustering images belonging to the same geo-cluster based on their visual features. The images are represented using a popular bag of visual words model (BoW) based on scale-invariant feature transform (SIFT) descriptors [51]. First, a certain number of keypoints are detected and described using the SIFT detector and descriptor. Further, k-means clustering is used to cluster the descriptors extracted from all images of a certain geographic area into 500 clusters (visual words). Finally, an image is represented with a 500-bin histogram, where each bin corresponds to a visual word in the codebook. In the following step, we cluster images from a particular geo-cluster into a certain number of visual clusters. For that, we again utilize the affinity propagation clustering using as the input image visual similarities

$$\mathbf{S}_v(i, j) = \mathrm{sim}(\mathbf{f}_i, \mathbf{f}_j) = e^{-||\mathbf{f}_i - \mathbf{f}_j||^2} \tag{4.2}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are the BoW feature vectors (histograms) of images i and j.

We conjecture that a frequency of appearance of an object or event in the images throughout the collection indicates its importance for the visual summary. Therefore, given the detected geo-visual clusters $C_l, l = 1 \dots k$, for an image i from the cluster C_l, we define the first component of the feature vector $\mathbf{x}_i$ of image i as $x_{i1} = |C_l|/N$, where N is the total number of images per location (here set to 100, as explained in Section 4.4.1).

4.5.2 Image popularity

In photo sharing websites such as e.g., Flickr, image view count and number of comments are generally believed to be correlated, at least weakly, with the user-perceived image aesthetic appeal. As such information is usually relatively easy to obtain and does not imply additional computational costs, without going into a deeper analysis of the factors that influence popularity of social media, we decided to include it in our image representation.

View Count: An image is represented by its view count (x_{i2}) as well as the mean and variance of the view counts of images in the same geo-visual cluster (x_{i3} and x_{i4}).

Number of Comments: Number of comments posted on an image together with the mean and variance of the number of comments associated with the images belonging to the same cluster are added as x_{i5}, x_{i6} and x_{i7}.

4.5.3 Image aesthetic appeal

To model image aesthetic appeal we make use of proven and computationally inexpensive aesthetic appeal indicators, i.e., image aspect ratio, colorfulness, luminance and sharpness.

Aspect Ratio: Our user study indicates that the users have a strong preference towards "landscape" image orientation or in other words the images having larger width than height. The exceptions are e.g., images of a particularly tall building such as Eiffel Tower (c.f. Fig. 4.3). We compute the aspect ratio as $x_{i8} = w/h$, where w and h are the image width and height. Additionally, we represent an image with the mean and variance of the aspect ratio of all images from the same geo-visual cluster (x_{i9} and x_{i10}).

Colorfulness: Image colorfulness is evaluated using a metric proposed in [26], which shows a high correlation with human perception. Then, an image i is represented with its estimated colorfulness (x_{i11}) as well as the mean and variance of the colorfulness of the images belonging to the same geo-visual cluster (x_{i12} and x_{i13}).

Luminance: To calculate the global luminance of an image, we first convert it from the RGB to YCbCr color space and then compute the mean value of the Y-channel in all pixels. The image i is represented with its luminance (x_{i14}) as well as the mean and variance of the luminance of all images belonging to the same geo-visual cluster (x_{i15} and x_{i16}).

Sharpness: Image sharpness is evaluated using the publicly available software [58], which computes the cumulative probability of blur detection (CPBD) at the edges in the image [57]. Similar to colorfulness and luminance, we represent each image with its estimated sharpness (x_{i17}) as well as the mean and variance of sharpness of semantically related images from the same geo-visual cluster (x_{i18} and x_{i19}).

4.5.4 Sentiment analysis

Compared to some other content sharing websites, such as e.g., YouTube, Flickr images are associated with a smaller average number of comments, which are often not very polarized. While in YouTube a controversial semantic theme of a video might cause an intensive discussion amongst visitors, such behavior is less frequently observed in Flickr. Still, as recently suggested in [93], Flickr comments might carry a valuable information for estimating sentiment of an image.

Since Flickr comments are often written in different languages, we first translate them all into English using Google Translate service. Further, for the terms appearing in the Whissell's Dictionary of Affect in Language (DAL) we obtain the valence, activation and imagery scores. Valence value indicates the level of pleasantness or unpleasantness that a particular word expresses, activation indicates the associated arousal level and the imagery designates whether a particular word is easy or hard to imagine. For example, the word *beautiful* is associated with a maximum valence value 3, while the word *terrible* has the lowest valence of 1. Contrary to the word *love*, associated with a relatively high activation of 2.6, the word *scenery* has an activation of only 1.2. Finally, an example of the word with the lowest imagery of 1 is *like*, while the words designating objects, such as e.g., *camera* or *house* are associated with a high imagery value of 3. Although in e.g., narrative peak detection scenarios [37] usually only valence and activation are utilized, we conjecture that even imagery could provide a potentially valuable information for determining sentiment of a comment. For example, a high imagery of the words in the comments on a Flickr image might indicate an absence of feedback containing strong sentiments or rather descriptive nature of the comments.

Although, in general, natural language processing (NLP) may prove beneficial for the sentiment analysis, here we choose not to perform it for several reasons. Namely, as already mentioned earlier in this section, the Flickr comments are relatively short, seldom polarized and frequently express appreciation of the image, which simplifies the sentiment analysis and reduces the need for NLP. Additionally, since the sentiment analysis is not the main focus of this chapter, we choose to perform it in a simple and computationally inexpensive manner that was proven effective in related work such as e.g., [18].

We compute the mean valence, activation and imagery values for the words in a comment and then average it over all comments posted on that image (feature vector components x_{i20}, x_{i21} and x_{i22}). Finally, we also represent an image with the mean and variance of valence (x_{i23} and x_{i24}), activation (x_{i25} and x_{i26}) and imagery (x_{i27} and x_{i28}) features across images belonging to the same geo-visual cluster.

4.6 User-informed image selection

To facilitate the selection of images for the summary we set as our target to produce a ranked list of images per location, where the rank position of an image serves as an indicator of its suitability for the visual summary. We approach learning to generate the ranked list in a user-informed fashion, first by selecting the training images from the human-created reference summaries and then by learning the ranking function taking the features from the previous section as the input.

We start the training data selection by sorting the images per location (collection subset consisting of 100 images) according to the number of MTurk workers that selected them for their summaries. Further, we choose a set of image preference pairs, $(i, j) \in \mathcal{P}$, each consisting of a top ranked and a bottom ranked image. In the selected set of preference pairs $\mathcal{P}$, a top or bottom-ranked image i can appear in only one preference pair and for each preference pair $(i, j) \in \mathcal{P}$, image i is preferred over image j. Then, to learn the ranking model, a well-known RankSVM method [36] could be used. In the method originally proposed by Joachims in [36], the RankSVM model is based on minimizing the following objective function

$$\frac{1}{2}\left\|\mathbf{w}\right\|^2 + C \sum_{(i,j)\in\mathcal{P}} \ell\left(\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{x}_j\right) \tag{4.3}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors representing images i and j, respectively, C is a regularization parameter and ℓ is a loss function, such as e.g., $\ell(z) = \max(0, 1 - z)$ in case of SVMLight implementation [36]. However, due to the relatively high computational costs associated with training of SVMLight, here we make use of a fast RankSVM method described in [11], whose clear notation we adopt in (4.3). The method is based on Newton optimization and avoids explicit computing of all possible difference vectors $\mathbf{x}_i - \mathbf{x}_j$ to significantly reduce the RankSVM training time.

As described in Section 4.4.1, the locations in Paris at which the images were captured are often rather different in terms of both semantic density and visual homogeneity. We conjecture that the images selected for the visual summary by an MTurk worker must be considered in the context of images of that particular geographic area. For example, their diversity and representativeness strongly depends on e.g., the diversity of the starting image set, whether the objects and events depicted in the images are perceived as significant or important etc. Also, an image might be selected not because it is particularly appealing, but simply because most of the other images are perceived as unappealing. Therefore, we train RankSVM separately for each of t locations (collection subsets) in the training set. Given a test image set, we apply the trained models and produce t lists of images ranked according to their suitability for the visual summary. Finally, a rank aggregation algorithm is applied to produce the final image ranking.

4.7 The pyramid approach to set evaluation

As discussed in Section 4.3.2, a common problem in evaluation of e.g., document summaries and machine translations is a low inter-user agreement (e.g., [60]). Fig. 4.5a shows a histogram of the level of agreement between summaries manually produced by the MTurk workers. The histogram indicates that the agreement is in general very low, with the mean of 1.5 and median of 1. In other words, two reference summaries usually have only one image in common, which makes the evaluation algorithms such as e.g., BLEU [71], ROUGE [49] and VERT [48] practically inapplicable. However, we also observe a high inter-user agreement in case of some images. We conjecture that those images, frequently appearing in the reference summaries are indeed the most important for the visual summary. The histogram in Fig. 4.5b shows in how many locations the workers agree on the most popular image, where the image is considered as the most popular if it appears in the largest number of reference summaries.

We observe that in each collection subset, there is at least one image that has been selected for the visual summary by at least 6 workers and that the median agreement on the most popular image per location is 9. To optimally exploit the inter-user agreement, we follow the idea of [60] and propose a pyramid approach for evaluating the suitability of images for the visual summary. As illustrated in Fig. 4.6, each pyramid tier consists of the images appearing in the same number of visual summaries. The most frequently selected images are placed in the top tier, while the bottom tier is composed of images that were selected by a single MTurk worker only. Images that do not appear in any of 20 reference summaries generated for a given location are considered unimportant and therefore discarded.

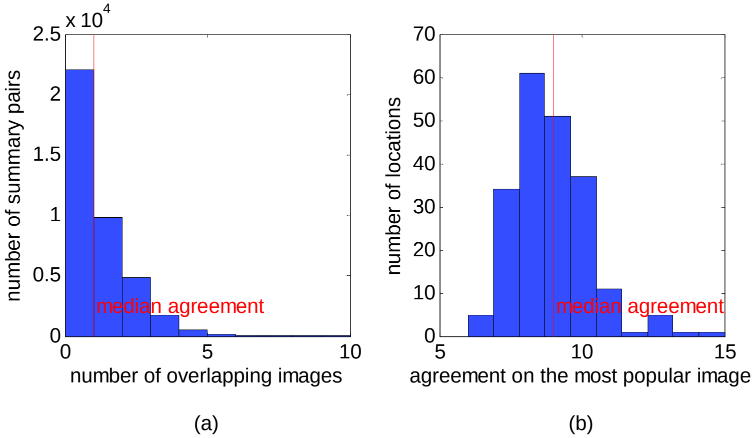


Figure 4.5: Illustration of the agreement between human-made reference summaries; (a) Histogram of the size of overlap between reference summaries produced by different workers, which shows the number of summary pairs having a particular number of images in common. (b) Histogram of the level of agreement on the most popular image, which shows the number of locations for which a particular number of workers selected the most popular image for their summary. As the most popular image in a given location (collection subset), we consider an image that appears in the largest number of summaries.

For example, in the particular case of location for which an illustration is shown in Fig. 4.6, the pyramid has 9 tiers and the image in the top tier appears in 11 out of 20 reference summaries.

We conjecture that an *optimal* set should include all images from the upper tiers and draw the remaining images from the last tier needed to reach a specified set size. In case of pyramid depicted in Fig. 4.6, an optimal 5-image set should include all images from the tiers T_n and T_{n-1} as well as 2 images from the tier T_{n-2}. Obviously, several optimal sets can be created as described above and in this particular example the number of such optimal sets is 3. According to the pyramid approach an optimal set $\tilde{R}$ with N_R images would receive the maximum score $d_{\max}$ computed as follows

$$\begin{aligned} d_{\max} &= \sum_{i=\theta+1}^{n} i \times |T_i| + \theta \times \left(N_R - \sum_{i=\theta+1}^{n} |T_i| \right), \\ \theta &= \max_{i} \left(\sum_{j=i}^{n} |T_j| \geq N_R \right) \end{aligned} \tag{4.4}$$

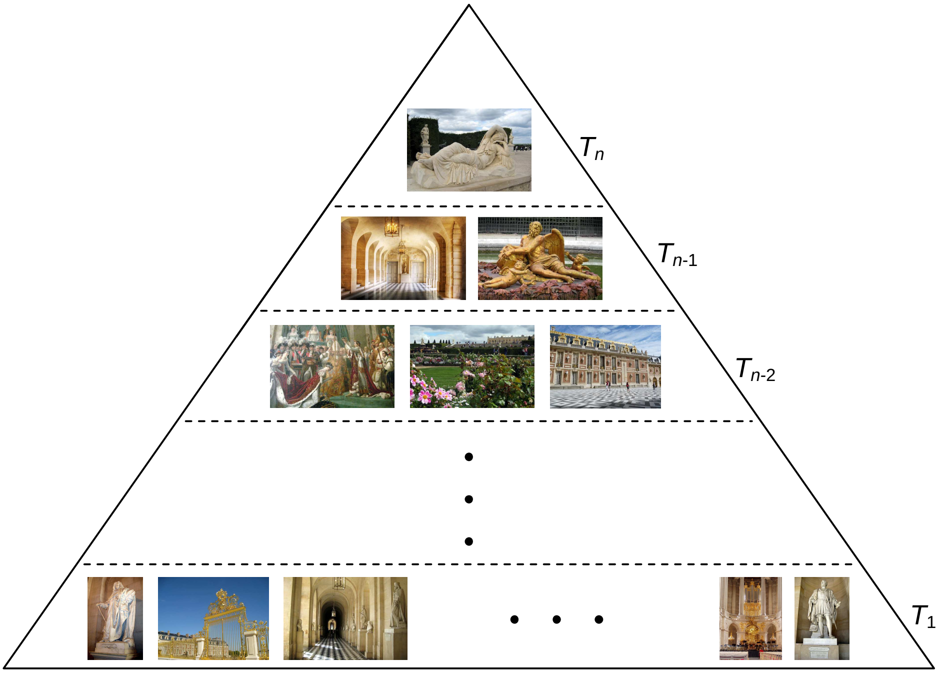


Figure 4.6: Illustration of the pyramid structure, where each tier consists of the images appearing in the same number of reference summaries. Each image in the bottom tier T_1 appears in only one reference summary, while the images in the top tier T_n are those most frequently selected for the visual summary.

Then, an arbitrary set R with N_R images receives the score d

$$d = \frac{1}{d_{\max}} \times \sum_{i=1}^{n} i \times |T_i \cap R| \tag{4.5}$$

For example, as the pyramid depicted in Fig. 4.6 has 9 tiers, the optimal 5-image set would receive the maximum score $d_{\max} = 1 \times 9 + 2 \times 8 + 2 \times 7 = 39$. As a side note we would like to mention that of all VERT variants, VERT-R1 bears the closest resemblance to the proposed evaluation metric.

In Section 4.9.1 we will demonstrate that the pyramid score is indeed effective in evaluating the quality of an image set.

4.8 Experimental setup

4.8.1 Baselines used for evaluation of the pyramid score

In Section 4.9.1 the effectiveness of the pyramid score is evaluated through comparison of the values obtained for reference summaries and the summaries composed of either the least popular images or the summaries output by the approaches that do not take into account image popularity, aesthetic appeal or sentiment. More particularly, for comparison we use the following baselines

Low View Count: The images with the lowest view count are selected.

RWR-RD [81]: The approach utilizes random walk with restarts over a multi-layer graph modeling text associated with the images, visual features extracted from them as well as the information about users and their social network to select a set of representative and diverse images of a particular geographic area. The approach is designed such to show various aspects of the area, but it is unaware of image popularity, aesthetic appeal or sentiment.

MA Clustering: The approach is based on the same multi-layer graph [81] as the RWR-RD approach described above and utilizes random walk with restarts algorithm to compute multimodal image similarities. The images are further clustered using the affinity propagation clustering [22] based on the computed similarities and the cluster centroids are selected for the result image set. Like RWR-RD, the approach does not focus on aesthetic properties of the images and their popularity.

Ensemble Clustering: The images are first clustered independently using the low-level visual features and the text associated with them [81] and then the ensemble clustering approach [100] is applied to produce a single, reinforced clustering. Finally, the clusters' visual centroids are selected for the visual summary of a collection. The approach does not make use of information about image popularity, aesthetics or sentiment.

4.8.2 Baselines used for image selection evaluation

In Section 4.9.2 we evaluate our proposed image selection approach by means of the pyramid score and compare it with two intuitive control baselines (Random and High VC) as well as the proven visual summarization approaches (MAC-VC and EC-VC).

Random: Images are randomly sampled from the collection. We find it important to report the performance of a random baseline in scenarios such as the one described in this chapter, to investigate whether the performance of the tested approaches differs significantly from random.

High VC: Images are selected based on a high view count. Although view count in general might be considered as an unreliable popularity indicator due to e.g., ease of manipulation and bias towards highly popular content causing the long tail problem [73], it is usually considered to be (weakly) correlated with the aesthetic appeal and sentiment.

MAC-VC: A modification of MA Clustering approach described in the previous section. Instead of choosing cluster centroids for the final results list, an image with the highest view count is selected to represent each cluster.

EC-VC: A variant of Ensemble Clustering approach described in the previous section, which, instead of choosing visual centroids, samples an image with the highest view count from each cluster for the final results list.

4.8.3 Training RankSVM and rank aggregation

As explained in Section 4.6, we train RankSVM model separately for all t locations in the training set and produce t ranked lists of images for a test location. We experimentally set the number of preference pairs $|\mathcal{P}| = 20$ (cf. Section 4.6) as a tradeoff between three factors - the number of training samples (preferably larger), the quality of samples (preferably only a small fraction of top and bottom ranked samples should be used) and the total number of images per collection subset (in this particular case - 100). Once the individual ranked lists are produced, the final ranking is generated through rank aggregation. In the past decade a number of approaches to rank aggregation have been proposed [19, 74]. In our exploratory experiments the approach proposed by Pihur et al. [74] yielded a good performance, but due to a high computational complexity and the fact that the main focus of this chapter are not the approaches for rank aggregation, we opted for a lightweight alternative. Here we perform the rank aggregation by simply computing the average rank of an image across all t lists. In our exploratory experiments such approach was proven to yield insignificantly lower performance than computationally intensive alternatives such as e.g., [74].

4.9 Experimental results

Through the experiments presented in this section we aim to answer the following research questions:

1. Is the pyramid score introduced in Section 4.7 effective in estimating the quality of an image set?
2. Does our proposed approach succeed in selecting a set of images suitable for visual summarization?
3. Is the performance well distributed across locations/ collection subsets?
4. Which features are the most important for isolating images with desired properties?
5. What is the relationship between different features?
6. Is our proposed approach applicable in case of image collections missing information richness of social media?

4.9.1 Evaluation of the pyramid score

We conjecture that a good evaluation metric should yield significantly higher scores for the reference summaries manually generated by the MTurk workers than for apparently lower-quality image sets or image sets automatically generated without taking into account sophisticated features, such as those related to e.g. image aesthetic appeal and sentiment. Our goal is also to investigate how the scores change with the varying number of reference summaries used to create a pyramid. Therefore, we vary the number of reference summaries used for pyramid building from 2 to 18 and compute the scores for the remaining reference summaries and three summarization approaches described in Section 4.8.1: LVC, RWR-RD, MAC and EC. The scores obtained for the reference summaries are simply averaged for easier comparison. All scores obtained for a particular approach under the same setting are averaged across all locations.

The graphs in Fig. 4.7 show that the computed scores generally grow with the increasing number of reference summaries used to construct the pyramid. Further, the scores averaged over remaining reference summaries are significantly higher than those computed for a set of images selected

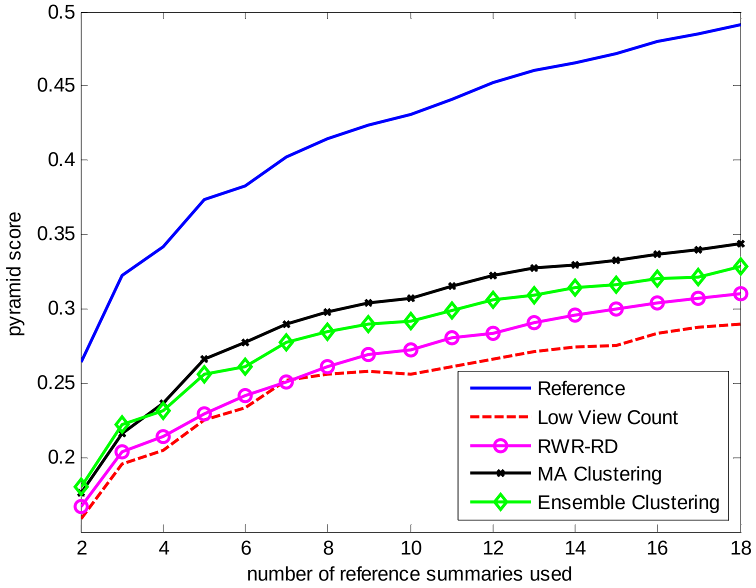


Figure 4.7: Variation of pyramid score depending on the number of reference summaries used for pyramid construction. The scores are computed for the remaining reference summaries and the four visual summarization approaches.

based on a low view count and the baselines that do not take into account image aesthetic appeal and sentiment.

In Fig. 4.8 we show for which percentage of locations image set produced in a particular way yields the highest score. This percentage increases with the increasing number of summaries used to construct the pyramid. Again, the pyramid score appears to be effective in discriminating between the high quality image sets manually created by the MTurk workers and those created automatically.

4.9.2 Evaluation of the proposed image selection approach

Here we compare the performance of our proposed approach for user-informed image selection with the performance of several competitive baselines described in Section 4.8.2: Random, High VC, MAC-VC and EC-VC. As the figures 4.7 and 4.8 indicate that the margin between scores computed for different approaches increases with the increasing number of reference summaries, for pyramid construction we make use of all 20 manually created summaries. We opt for a “leave-one-out” strategy simultaneously

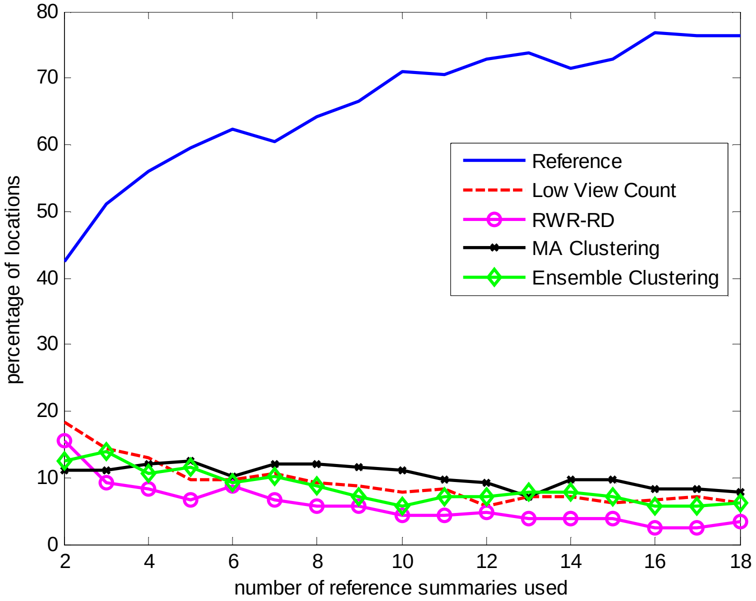


Figure 4.8: Comparison of the pyramid scores assigned to the reference summaries and the four visual summarization approaches. The percentage of locations for which a particular approach yields the highest score is reported.

training RankSVM on $t = 206$ collection subsets and apply the trained model on the remaining subset (location). Finally, for easier comparison we report the scores averaged over all 207 locations.

The performance comparison of our RSVM-CAS selection approach and the four baselines in terms of pyramid score averaged over all 207 locations is presented in Table 4.1. Our proposed approach clearly selects higher-quality image sets of various sizes N_R. Further, although Random image selection yields a reasonable collection sampling in terms of e.g., representativeness and diversity [81], this approach does not take into account criteria found important by the users when creating visual summaries, such as e.g., image aesthetic appeal and sentiment. Further, view count shows a high correlation with the user-perceived image aesthetic appeal and might be considered as a solid selection strategy in cases when a low computational complexity is required. However, view count alone is often seen as an unreliable popularity indicator as it can be unavailable and manipulated, but it can also lead to a bias towards the mainstream content. Although our proposed RSVM-CAS approach makes use of view count and number

Table 4.1: Performance of our RSVM-CAS selection approach and the four baselines reported in terms of pyramid score averaged over all 207 locations

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
Random	0.253	0.256	0.301	0.350
High VC	0.414	0.450	0.478	0.502
MAC-VC	0.345	0.388	0.432	0.457
EC-VC	0.362	0.410	0.442	0.464
RSVM-CAS	**0.566**	**0.574**	**0.596**	**0.622**

Table 4.2: Performance of our RSVM-CAS selection approach and the four baselines reported in terms of percentage of locations for which a particular approach is the best performer

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
Random	2.4	0.5	0.5	0.5
High VC	16.9	18.8	15.5	15.9
MAC-VC	10.6	4.3	7.7	5.4
EC-VC	8.3	5.8	9.6	7.7
RSVM-CAS	**61.8**	**70.6**	**66.7**	**70.5**

of comments, we conjecture that the other features modeling image aesthetic appeal, sentiment and context make it more robust to those and similar negative factors.

4.9.3 Performance distribution across image collection

To investigate whether the performance of our proposed RSVM-CAS approach is well distributed across the collection, we compute the percentage of locations for which a particular approach performs better then the alternatives. As shown in Table 4.2, our proposed RSVM-CAS approach is the best performer in largest number of locations for various sizes N_R of the output image set.

4.9.4 Analysis of feature discriminativeness

Here we compare the effectiveness of each feature used in discriminating between images that appear frequently in the reference summaries and

Table 4.3: Ranked list of features sorted by their effectiveness in discriminating between images appearing most frequently in the reference summaries and those selected least frequently

Rank	Feature	Rank	Feature
1	aspect ratio (x_{i8})	15	mean sharpness (x_{i18})
2	mean aspect ratio (x_{i9})	16	luminance (x_{i14})
3	colorfulness (x_{i11})	17	var view count (x_{i4})
4	view count (x_{i2})	18	sharpness (x_{i17})
5	nr comments (x_{i5})	19	cluster size (x_{i1})
6	valence (x_{i20})	20	mean luminance (x_{i15})
7	activation (x_{i21})	21	var nr comments (x_{i7})
8	mean view count (x_{i3})	22	var imagery (x_{i28})
9	imagery (x_{i22})	23	var valence (x_{i24})
10	mean colorfulness (x_{i12})	24	var activation (x_{i26})
11	mean activation (x_{i25})	25	var sharpness (x_{i19})
12	mean valence (x_{i23})	26	var colorfulness (x_{i13})
13	mean nr comments (x_{i6})	27	var aspect ratio (x_{i10})
14	mean imagery (x_{i27})	28	var luminance (x_{i16})

the least popular ones. For each location we select 20 images appearing most frequently in the reference summaries and treat them as the positive class. Similarly, for the negative class we select 20 images that appear least frequently in the reference summaries. Further, we perform the forward feature selection for classification using the 1-Nearest Neighbor error criterion, which first selects a single most discriminative feature and which further iteratively selects the feature that improves most the discriminativeness of the feature set. Once the list of features sorted according to their discriminativeness is produced for each location, we perform the rank aggregation by averaging the rank of each feature across all 207 ranked lists. The ranked list of features is shown in Table 4.3.

Surprisingly, image aspect ratio and colorfulness features emerge as the most discriminative, which further confirms our assumption that the users are to a large extent driven by image aesthetic appeal when selecting images for the visual summary. For example, the most frequently occurring aspect ratios (w/h) in the entire image collection are 1.2723, 1.4164, 0.7300, 1.4085, 0.6521 and 0.9540, while the most frequent aspect ratios amongst the images selected by the MTurk workers are 1.3333, 1.5015, 0.7500, 1.4970, 0.6660 and 1.0000. Our findings are in line with the outcomes of a user study discussed in [50], which show that, contrary to a common belief, the "golden ratio" (i.e., $w/h = 1.618$) may not be the most appealing image aspect ratio. Further, a high discriminativeness of mean aspect ratio feature confirms our assumption about importance of image context. We conjecture that in the case of e.g., "panoramic" spots many images will have a similar aspect ratio that best captures the content of the scene. In that sense, a similar aspect ratio of the images in a particular geo-visual cluster might be (implicitly) indicative of e.g., interestingness or visual appeal of a view from that location. Lightweight popularity indicators such as e.g., view count and number of comments are also positioned high in the ranked list. Finally, sentiment features extracted from image comments, namely valence, activation and imagery fall into the group of the most discriminative features as well. Valence and activation appear to be more important than imagery, which is not surprising, since those features provide more explicit information about sentiment of a word.

On the other hand, sharpness and luminance appear to be less important than the other aesthetic appeal and sentiment features. Also, a relatively low rank of cluster size feature might indicate that aesthetic attributes of the image as well as the sentiment it evokes play a more important role than the representativeness and diversity. Finally, when considering contextual features (i.e., mean and variance of a particular feature computed for semantically similar images, e.g., those in the same geo-visual cluster), mean is to be preferred to variance.

4.9.5 Relationship between different features

To complement the experiment from the previous section, here we investigate the correlation between different features. The heat map in Fig. 4.9 visualizes the relationship between features expressed in terms of median correlation coefficient computed over all 207 locations.

As shown in Fig. 4.9, there is no apparent correlation between view count and the image sentiment features - valence, activation and imagery. Also, image aesthetic appeal features including image aspect ratio and colorfulness, which emerged as the most discriminative features in the previous section, seem to be uncorrelated with the view count and number of comments. However, the number of comments shows a certain degree of correlation with the valence, activation and imagery, which is somewhat

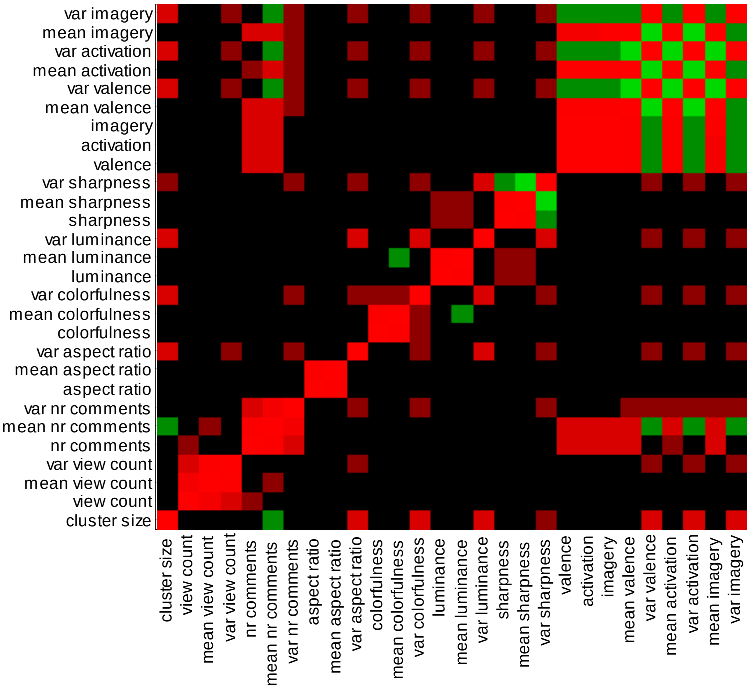


Figure 4.9: Relationship between features expressed in terms of median correlation coefficient computed over all locations. Red and green colors indicate positive and negative correlation.

expected considering the fact that those features were extracted from the image comments. Finally, we observe a high correlation between valence, activation and imagery features.

4.9.6 Extension to non-annotated image collections

Compared to rich social media, offline collections are often poorly (if at all) annotated and images are lacking the useful information such as e.g., title, description, tags, comments and view count. Here we investigate the effectiveness of our approach in such cases when only information automatically captured by the camera is available, i.e. image content and automatically captured geo-coordinates. Although the geo-tags available in Flickr are sometimes manually inserted by the users, for the purpose of this experiment we consider them all to be automatically generated by the capturing device. We conjecture that the increasing availability of capturing devices

Table 4.4: Performance of our RSVM-CA and RSVM-CAS selection approaches reported in terms of pyramid score averaged over all 207 locations

Selection Method	$N_R = 5$	$N_R = 10$	$N_R = 15$	$N_R = 20$
RSVM-CA	0.529	0.534	0.560	0.578
RSVM-CAS	**0.566**	**0.574**	**0.596**	**0.622**

(e.g., cameras and smart phones) with a high positioning accuracy, make the scenario realistic. In the cases when the geo-coordinates are not available at all, a clustering described in Section 4.5.1 could be performed based on e.g., visual features only.

Following the scenario described above, we retain only the following features: cluster size (x_{i1}), aspect ratio (x_{i8}), mean aspect ratio (x_{i9}), var aspect ratio (x_{i10}), colorfulness (x_{i11}), mean colorfulness (x_{i12}), var colorfulness (x_{i13}), luminance (x_{i14}), mean luminance (x_{i15}), var luminance (x_{i16}), sharpness (x_{i17}), mean sharpness (x_{i18}) and var sharpness (x_{i19}). Performance of our proposed approach utilizing contextual and image aesthetic appeal features only (RSVM-CA) is shown in Table 4.4.

Comparing the results in Table 4.4 with those in Table 4.1 we observe that the RSVM-CA manages to outperform the approaches utilizing rich information available in social media, while being agnostic to image aesthetic appeal and image sentiment. However, the performance drop compared to RSVM-CAS confirms the importance of popularity indicators and image sentiment features.

4.10 Discussion and future work

We have used information about how humans select images for visual summaries, which was collected with a large-scale crowdsourcing study, as the basis for a novel method for automatically selecting images for visual summarization. The crowdsourcing study revealed inherent properties of images that are important for humans and also provided us with training data. Our approach uses features based on these properties and RankSVM method to generate a list of images ranked by their suitability for inclusion in a visual summary. As such, the selected image set can be used as a “general purpose” visual summary or as a starting point in building a summary with particular properties.

We discuss a phenomenon of a low inter-user agreement and prove effectiveness of the metric based on the pyramid score in evaluating the quality of a selected set of images. Both the evaluation metric and our image selection approach are tested on a collection of geo-referenced Flickr images. Under various conditions our approach has proven effective in generating image sets composed of images that are frequently selected for the visual summaries by humans. The approach shows a potential for use in both information-rich social media environments as well as in the case of non-annotated image collections.

Both our large-scale user study and the analysis of feature discriminativeness indicate the effectiveness of the computationally inexpensive image aesthetic appeal features. Our analysis places image popularity indicators and sentiment features in the group of most discriminative features and their use brings an additional improvement in the system. Surprisingly, no apparent correlation has been found between image aesthetic appeal features and the popularity indicators, which might indicate that some other properties have a larger impact on the popularity of social media. We leave a deeper study of the relations between different features for the future work.

Although we prove the effectiveness of the pyramid score in evaluating the quality of selected image sets from the aspect of selection of images found by users as suitable for visual summarization, it does not explicitly evaluate attributes such as e.g., image set diversity. We believe that the largest potential for better incorporating diversity into the evaluation metric is a more sophisticated means of determining the semantic similarity between images. Further, we demonstrate that the amount of data and a low inter-user agreement deem the traditional evaluation metrics such as e.g., ROUGE and BLEU practically inapplicable, creating a need for the metrics taking into account specificities of multimedia content. Also, as we show in this chapter, the way the users perceive the image selection criteria and their interplay are often more complex than the related work in the field often suggests. In our future work we will further investigate those criteria and the means to evaluate them.

Currently we are estimating sentiment of the comments posted in response to the images only, but we plan to investigate whether useful affective information might be extracted from image title, tags and description generated by the uploader. Finally, our future work will also include a deeper analysis of the factors that influence effectiveness of our approach.

Chapter 5

Reflections and recommendations

This thesis presents the results of our research on advancing the relevance criteria for video search and visual summarization. To address the problem from a broader perspective, we considered the practical use-cases associated with two radically different settings, namely a professional, unlabeled video collection and information-rich social media. In sections 5.1 and 5.2 we briefly summarize the achievements reported in chapters 2, 3 and 4 with regard to video search and visual summarization, respectively, provide the answers to the related research questions formulated in Chapter 1, reflect on the encountered problems and give recommendations regarding future research. Section 5.3 concludes the thesis with some final remarks.

5.1 Video search in a professional collection setting

5.1.1 What has been achieved

In **Chapter 2** we addressed the problem of video search at the level of semantic theme in the setting of an unlabeled professional video collection. The objective was to provide an answer to the question of *how to facilitate video search at the level of semantic theme by relying on the visual and spoken content of the video only.*

To pursue an answer to this question, we developed a framework that is based on the query performance prediction principle and that aims at selecting the best out of the query modification and retrieval methods we proposed and tested for responding to a given topical query. An evaluation of this framework indicated that even in the complete absence of annotations, both the spoken content and the visual channel are useful for deriving the general subject matter of the video, provided that the right type of information from these channels is extracted. Joint exploitation of these channels was proven effective, to more than compensate their individual imperfections and to lead to more reliable predictions and, consequently, to improved retrieval performance compared to working with the visual and spoken (text) channels individually.

In particular, despite the rather modest performance of individual visual concept detectors in a general case, our proposed video representation, based on aggregating shot-level visual concept detections across the entire video, was demonstrated effective in capturing video similarities at the level of semantic theme. We showed that selecting a compact subset of the most discriminative visual concepts generally leads to a further improvement of retrieval performance and proposed to that end an unsupervised concept selection approach. Additional benefit of the concept selection step is that it makes our video representation relatively independent of the semantic coverage of the original concept set. Spoken content channel of the video is, in general, expected to carry a potentially more useful information about the semantic theme than the visual channel. Interestingly, concept-based indicators of query performance yielded a comparable performance to the state-of-the-art alternatives utilizing spoken content of the videos (i.e., text-based indicators).

5.1.2 Reflection and future work

Although our proposed retrieval framework generally improves the retrieval performance, there is still a large room for further improvement. Namely, both text-based and concept-based indicators of query performance appear to be only moderately successful in selecting the optimal retrieval setting for a given topical query. Additionally, they were shown to be sensitive to parameter setting, which implies a need for parameter re-optimization when switching to a new dataset. However, instead of attempting to optimize the performance of individual detectors, investigation into the methods for their fusion may appear to be a more productive research avenue.

Furthermore, the finding that the visual content of the videos can be effectively utilized for their automatic comparison at the level of semantic theme has a high scientific merit, but the high computational costs associated with visual concept detection and its noisiness make a large-scale application in commercial search engines challenging at the moment. A possible step forward in addressing this problem may be identifying a compact set of the most discriminative visual concepts and then concentrating on significantly improving their performance.

Finally, our experience suggests that the manually generated metadata are generally more beneficial for video retrieval at a higher semantic level than the information automatically extracted from the video [44, 87]. More specifically, even the straightforward approaches relying on e.g. manually generated metadata and the information about user interactions with the content frequently yield significantly better results than the sophisticated alternatives utilizing only the information automatically extracted from the content. Therefore, the approaches for facilitating a more efficient annotation in the professional archives by e.g., bringing them closer to the social communities or by making the use of crowdsourcing, are certainly worth investigating.

5.2 Visual summarization in a social media setting

5.2.1 What has been achieved

We approached the problem of visual summarization by first addressing the question of *how to maximize the quality of a visual summary, given the available social media information resources.* In this respect, and taking a realistic application as a test bed, we proposed in **Chapter 3** a novel approach that makes use of representative, but diverse community contributed images, to create a visual summary showing all relevant aspects of a geographic area. In addition, we proposed a novel protocol for a cost and time efficient evaluation of such created visual summaries, which makes use of metadata associated with the images only and does not require an additional input of the human assessors. We demonstrated the feasibility of fusing visual, text and user modalities via a multi-layer graph for a more reliable measuring of image similarities that serve as the basis for an iterative summary construction. We proved that such fusing of heterogeneous modalities available in a social media setting, enables capturing of explicit and implicit relations between images defined at a higher semantic level.

Chapter 3 showed that the proposed modality-dependent weighting applied to the edges of the graph brings an additional performance improvement. Analysis of the unimodal algorithm alternatives, based on the graphs in which only a single modality is kept, reveals that, surprisingly, the user modality contributes the most to selecting representative and diverse images. This may be explained by the observation that different users have different interests and therefore capture, to a certain degree, authentic views of a geographic area. In addition, as the social subgraph captures e.g., users' common interests in particular (sets of) images, image similarities derived from it are inherently at a higher semantic level. Our iterative approach to selecting representative and diverse images outperforms the common state of the art alternatives. Compared to the alternatives, our summarization algorithm is less biased towards popular images and therefore better addresses the “long tail” problem.

In the next step, and building on the experience from Chapter 3, we addressed in **Chapter 4** another critical research question related to visual summarization, namely *whether it is possible to automatically identify the images that a user would consider suitable for creating a visual summary.* To get an insight into how humans perform visual summarization, we run a crowdsourcing experiment and collected a large number of human created visual summaries as well as the justifications for image inclusion in the summary. Based on the outcomes of the user study, we proposed a heterogeneous image representation based on image content, context, popularity, aesthetic appeal and sentiment. Using the human created visual summaries as the ground truth, we deployed the proposed feature representation to learn to rank images according to the likelihood that they would be included in the summary by the humans. In addition, we addressed a challenging problem of the automatic evaluation of visual summaries, which has gained relatively little attention so far, and proposed an evaluation protocol inspired by the metrics used for assessment of text summaries. Finally, we performed an analysis of the correlation between different features used to represent the images.

The most general message of Chapter 4 is that it is indeed possible to learn to identify images that the humans would find suitable for visual summarization. Crowdsourcing shows a great promise in understanding the users' information needs and here it aids the feature design, learning and evaluation phase. We demonstrated that analyzing the image content alone, which is the case with most state of the art visual summarization algorithms, yields sub-optimal results and therefore, image properties “orthogonal to the content”, such as e.g., aesthetic appeal, popularity and sentiment should be taken into account as well. In fact, we showed that some of those features are indeed the most useful for discriminating between images based on the frequency of appearance in human-created visual summaries. The proposed evaluation metrics was shown effective in mimicking human judgment about the quality of the selected image set. Finally, we demonstrated that our user-informed image selection approach could be applied not only to information-rich social media, but to e.g., unlabeled image collections as well.

5.2.2 Reflection and future work

The graph-based approaches, including the approach presented in Chapter 3, show promising results in modeling heterogeneous information associated with the images in a social media setting and in computing multimodal image similarities. However, the richness of the information derived from a graph and associated with the images leads to a high dimensionality of the adjacency matrix of the graph and therefore to high computational costs and reduced scalability. While the computational complexity was not the focus in Chapter 3, we see searching for possibilities to reduce the computational complexity as the next critical step in enabling effective practical deployment of the proposed methods. A promising way to approach this may be graph partitioning [6, 33].

The automatic evaluation of visual summaries based on the human-created reference summaries still remains a challenging problem. The evaluation protocol introduced in Chapter 4 was demonstrated effective in automatically evaluating the quality of the selected image set in the context of the problem addressed in the chapter, namely, selection of images likely to be included in the visual summary by the humans. However, the next logical question - how to compose a high-quality summary using the images suggested by our proposed method and evaluate it afterwards - remains largely open. An image set composed of good candidates for the visual summary would receive a high score according to the proposed metrics, but this does not automatically imply that this set would be a good visual summary according to human judgment. As a major obstacle to developing better summarization algorithms and metrics for evaluating the quality of visual summaries, we see the absence of means for reliably computing the semantic similarity between images, which is a necessary step in enforcing and evaluating the criteria determining the quality of a summary as a whole (e.g., semantic diversity).

5.3 Final remarks

Rapid increase in the amount of multimedia content produced in a professional setting and exchanged in the social media environments has made the task of finding a relevant information challenging. Like in many other research disciplines focusing on artificial intelligence problems, the biggest obstacle in designing the effective multimedia information retrieval solutions is the inability of machine to automatically extract the meaning from the content, which matches the level of human interpretation. While most proposed solutions operate at the lower semantic levels, where the multimedia items and their relations can be analysed more easily, the actual user information needs are often specified at a relatively high semantic level, which is where the focus of this thesis lies.

The results presented in the thesis are promising and indicate that the answer to the main research question of the thesis, namely, *whether video search and visual summarization can be performed based on the relevance criteria defined at a higher semantic level*, is positive. The requirement is, however, that the innovative ways of representing multimedia content are deployed, which reach beyond the conventional multimedia content analysis and rely whenever possible on the social information resources and the emerging technologies such as e.g., crowdsourcing, for a better understanding of the users' actual needs and the way they interpret the content in a given situation. When building such representations, no information resource should be discarded or advertised without a deeper consideration, because the optimal choice of the modalities and features to be used depends highly on the properties of the multimedia collection and the specified user information need.

The relevance criteria and their exact interplay appear to be significantly more complex than commonly assumed and while in the past technological limitations forced adoption of various simplifying assumptions, the unprecedented possibilities of collecting explicit and implicit information about users and their information needs have announced a breakthrough in designing relevance models as they are supposed to be.

Bibliography

[1] "Encyclopædia Britannica," http://www.britannica.com/.

[2] "Wikipedia," http://www.wikipedia.org/.

[3] "Yahoo! advertising solutions," http://advertising.yahoo.com/article/flickr.html, Assessed in October 2012.

[4] R. Aly, A. Doherty, D. Hiemstra, and A. Smeaton, "Beyond shot retrieval: Searching for broadcast news items using language models of concepts," in *Advances in Information Retrieval*, ser. LNCS. Springer Berlin / Heidelberg, 2010, vol. 5993, pp. 241–252.

[5] M. Ames and M. Naaman, "Why we tag: motivations for annotation in mobile and online media," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 971–980.

[6] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using pagerank vectors," in *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 475–486.

[7] D. Arijon, *Grammar of the Film Language.* Silman-James Press, 1976.

[8] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He, "Music recommendation by unified hypergraph: combining social media information and music content," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 391–400.

[9] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T. Huang, "A worldwide tourism recommendation system based on geotagged web photos," in *IEEE*

International Conference on Acoustics Speech and Signal Processing, ser. ICASSP '10, march 2010, pp. 2274 –2277.

[10] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 335–336.

[11] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Inf. Retr.*, vol. 13, no. 3, pp. 201–215, jun 2010.

[12] M. Clements, "Personalised access to social media," Ph.D. dissertation, TU Delft, Delft, The Netherlands, 2010.

[13] M. Clements, A. P. De Vries, and M. J. T. Reinders, "The task-dependent effect of tags and ratings on social media access," *ACM Trans. Inf. Syst.*, vol. 28, pp. 21:1–21:42, November 2010.

[14] J. Coutaz, "Multimedia and multimodal user interfaces: A taxonomy for software engineering research issues," in *East-West HCI'92*, August 1992, pp. 229–239.

[15] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," in *Proc. 25th annual int. ACM SIGIR conf. on Research and development in information retrieval*, ser. SIGIR '02. ACM, 2002, pp. 299–306.

[16] H. T. Dang, "Overview of DUC 2006," in *Proceedings of the Document Understanding Conference*, ser. DUC '06, 2006.

[17] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proceedings of the 9th European conference on Computer Vision - Volume Part III*, ser. ECCV'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 288–301.

[18] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis," in *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, april 2008, pp. 507 –512.

[19] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web*, ser. WWW '01. New York, NY, USA: ACM, 2001, pp. 613–622.

[20] C. Eickhoff and A. de Vries, "How crowdsourcable is your task?" ser. CSDM '11, 2011.

[21] A. Esuli and F. Sebastiani, “SentiWordNet: A publicly available lexical resource for opinion mining,” in *Proceedings of the 5th Conference on Language Resources and Evaluation*, ser. LREC ’06, 2006, pp. 417–422.

[22] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *Science*, vol. 315, pp. 972–976, 2007.

[23] A. Hanjalic and L.-Q. Xu, “Affective video content representation and modeling,” *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143 – 154, feb. 2005.

[24] Q. Hao, R. Cai, X.-J. Wang, J.-M. Yang, Y. Pang, and L. Zhang, “Generating location overviews with images and tags by mining user-generated travelogues,” in *Proceedings of the 17th ACM international conference on Multimedia*, ser. MM ’09. New York, NY, USA: ACM, 2009, pp. 801–804.

[25] D. Harman and P. Over, “The DUC summarization evaluations,” in *Proceedings of the second international conference on Human Language Technology Research*, ser. HLT ’02. Morgan Kaufmann Publishers Inc., 2002, pp. 44–51.

[26] D. Hasler and S. E. Süsstrunk, “Measuring colorfulness in natural images,” in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 5007, Jun. 2003, pp. 87–95.

[27] C. Hauff, V. Murdock, and R. Baeza-Yates, “Improved query difficulty prediction for the web,” in *Proc. 17th ACM conf. on Information and knowledge management*, ser. CIKM ’08. ACM, 2008, pp. 439–448.

[28] A. Hauptmann, M. Christel, and R. Yan, “Video retrieval based on semantic concepts,” *Proceedings of the IEEE*, vol. 96, no. 4, pp. 602–622, april 2008.

[29] A. Hauptmann, R. Yan, and W.-H. Lin, “How many high-level concepts will fill the semantic gap in news video retrieval?” in *Proc. 6th ACM int. conf. on Image and video retrieval*, ser. CIVR ’07, 2007, pp. 627–634.

[30] J. He, M. Larson, and M. de Rijke, “Using coherence-based measures to predict query difficulty,” in *Advances in Information Retrieval*, ser. LNCS. Springer Berlin / Heidelberg, 2008, vol. 4956, pp. 689–694.

[31] J. He, W. Weerkamp, M. Larson, and M. de Rijke, “An effective coherence measure to determine topical consistency in user-generated content,” *Int. J. Doc. Anal. Recognit.*, vol. 12, pp. 185–203, October 2009.

[32] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, “Video search reranking via information bottleneck principle,” in *Proc. 14th annual ACM int. conf. on Multimedia*, ser. MM ’06. ACM, 2006, pp. 35–44.

[33] X.-S. Hua, “Image and video tagging in the internet era,” in *2nd Summer School on Social Media Retrieval*, ser. S3MR ’11, Antalya, Turkey, 2011. [Online]. Available: http://videolectures.net/s3mr2011_hua_tagging/

[34] B. Huurnink, K. Hofmann, and M. de Rijke, “Assessing concept selection for video retrieval,” in *Proc. 1st ACM int. conf. on Multimedia information retrieval*, ser. MIR ’08. ACM, 2008, pp. 459–466.

[35] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo, “CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection,” Columbia University, ADVENT Technical Report #223-2008-1, 2008.

[36] T. Joachims, “Optimizing search engines using clickthrough data,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD ’02. New York, NY, USA: ACM, 2002, pp. 133–142.

[37] B. Jochems, M. Larson, R. Ordelman, R. Poppe, and K. P. Truong, “Towards affective state modeling in narrative and conversational settings,” in *Proceedings of Interspeech 2010*, September 2010, pp. 490–493.

[38] S. E. Johnson, P. Jourlin, K. S. Jones, and P. C. Woodland, “Spoken document retrieval for trec-8 at cambridge university,” in *Proc. TREC-8*, 2000, pp. 197–206.

[39] G. Kazai, “In search of quality in crowdsourcing for search engine evaluation,” in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6611, pp. 165–176.

[40] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, “To search or to label?: predicting the performance of search-based automatic image classifiers,” in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, ser. MIR ’06. New York, NY, USA: ACM, 2006, pp. 249–258.

[41] L. S. Kennedy and M. Naaman, “Generating diverse and representative image search results for landmarks,” in *Proceeding of the 17th international conference on World Wide Web*, ser. WWW ’08. New York, NY, USA: ACM, 2008, pp. 297–306.

[42] C. Kofler, L. Caballero, M. Menendez, V. Occhialini, and M. Larson, “Near2me: an authentic and personalized social media-based recommender for travel destinations,” in *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, ser. WSM ’11. New York, NY, USA: ACM, 2011, pp. 47–52.

[43] I. Konstas, V. Stathopoulos, and J. M. Jose, "On social networks and collaborative recommendation," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 195–202.

[44] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones, "Automatic tagging and geotagging in video collections and communities," in *Proc. 1st ACM Int. Conf. on Multimedia Retrieval*, ser. ICMR '11. ACM, 2011, pp. 51:1–51:8.

[45] K. S. Lee, W. B. Croft, and J. Allan, "A cluster-based resampling method for pseudo-relevance feedback," in *Proc. 31st annual int. ACM SIGIR conf. on Research and development in information retrieval*, ser. SIGIR '08. ACM, 2008, pp. 235–242.

[46] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, june 2012, pp. 1346 –1353.

[47] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, Feb. 2006.

[48] Y. Li and B. Merialdo, "VERT: automatic evaluation of video summaries," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 851–854.

[49] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop.* Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[50] M. Livio, *The golden ratio: The story of phi, the world's most astonishing number.* Broadway, 2008.

[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[52] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[53] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[54] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of the visual aesthetic appeal of consumer videos," in *Proceedings of the 11th European conference on Computer vision: Part V*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–14.

[55] F. Nack, C. Dorai, and S. Venkatesh, "Computational media aesthetics: finding meaning beautiful," *Multimedia, IEEE*, vol. 8, no. 4, pp. 10 –12, oct-dec 2001.

[56] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, pp. 86–91, July 2006.

[57] N. Narvekar and L. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *Image Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2678 –2683, sept. 2011.

[58] N. D. Narvekar and L. J. Karam, "CPBD sharpness metric software," http://ivulab.asu.edu/Quality/CPBD.

[59] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proc. 15th int. conf. on Multimedia*, ser. MM '07. ACM, 2007, pp. 991–1000.

[60] A. Nenkova and R. J. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *HLT-NAACL*, 2004, pp. 145–152.

[61] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 557–566.

[62] P. Obrador and N. Moroney, "Low level features for image appeal measurement," in *Proceedings SPIE, Image Quality and System Performance VI*, ser. IS&T/SPIE, vol. 7242, 2009, pp. 72 420T–1–72 420T–12.

[63] P. Over, G. Awad, J. Fiscus, B. Antonishek, and G. Qu, "TRECVID 2010 an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TRECVID Workshop*. NIST, 2010, pp. 1–34.

[64] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, and A. F. Smeaton, "Trecvid 2009 – goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2009*. NIST, USA, 2010.

[65] P. Over, A. F. Smeaton, and G. Awad, "The TRECVID 2008 BBC rushes summarization evaluation," in *Proceedings of the 2nd ACM TRECVID Video Summarization Workshop*, ser. TVS '08. New York, NY, USA: ACM, 2008, pp. 1–20.

[66] K. Owczarzak and H. T. Dang, "Overview of the TAC 2011 summarization track: Guided task and AESOP task," in *Proceedings of the Text Analysis Conference*, 2011.

[67] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120.

[68] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 653–658.

[69] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang, "Summarizing tourist destinations by mining user-generated travelogues and photos," *Comput. Vis. Image Underst.*, vol. 115, pp. 352–363, March 2011.

[70] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, August 2010.

[71] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[72] M. L. Paramita, M. Sanderson, and P. Clough, "Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009," in *Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments*, ser. CLEF'09. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 45–59.

[73] Y.-J. Park and A. Tuzhilin, "The long tail of recommender systems and how to leverage it," in *Proceedings of the 2008 ACM conference on Recommender systems*, ser. RecSys '08. New York, NY, USA: ACM, 2008, pp. 11–18.

[74] V. Pihur, S. Datta, and S. Datta, "Rankaggreg, an r package for weighted rank aggregation," *BMC Bioinformatics*, vol. 10, no. 1, p. 62, 2009.

[75] A. Popescu, G. Grefenstette, and P.-A. Moëllic, "Mining tourist information from user-supplied collections," in *Proceedings of the 18th ACM conference on Information and knowledge management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 1713–1716.

[76] M. Potthast, B. Stein, F. Loose, and S. Becker, "Information retrieval in the commentsphere," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 68:1–68:21, Sep. 2012.

[77] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Comput. Linguist.*, vol. 28, no. 4, pp. 399–408, Dec. 2002.

[78] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in mechanical turk," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '10. New York, NY, USA: ACM, 2010, pp. 2863–2872.

[79] S. Rudinac, M. Larson, and A. Hanjalic, "Exploiting visual reranking to improve pseudo-relevance feedback for spoken-content-based video retrieval," in *10th Workshop on Image Analysis for Multimedia Interactive Services*, ser. WIAMIS '09, 2009, pp. 17 –20.

[80] S. Rudinac, A. Hanjalic, and M. Larson, "Finding representative and diverse community contributed images to create visual summaries of geographic areas," in *Proceedings of the 19th ACM international conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 1109–1112.

[81] S. Rudinac, A. Hanjalic, and M. Larson, "Generating visual summaries of geographic areas using community contributed images," *IEEE Transactions on Multimedia*, 2013, IEEE Early Access Article.

[82] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *IEEE Transactions on Multimedia*, accepted for publication with mandatory minor revisions.

[83] S. Rudinac, M. Larson, and A. Hanjalic, "Visual resampling for pseudo-relevance feedback during speech-based video retrieval," in *Proceedings of the Information Retrieval 2009 Workshop at LWA 2009*, Darmstadt, Germany, 2009.

[84] S. Rudinac, M. Larson, and A. Hanjalic, "Exploiting noisy visual concept detection to improve spoken content based video retrieval," in *Proc. ACM int. conf. on Multimedia*, ser. MM '10. ACM, 2010, pp. 727–730.

[85] S. Rudinac, M. Larson, and A. Hanjalic, "Exploiting result consistency to select query expansions for spoken content retrieval," in *Advances in Information Retrieval*, ser. LNCS. Springer Berlin / Heidelberg, 2010, vol. 5993, pp. 645–648.

[86] S. Rudinac, M. Larson, and A. Hanjalic, “Visual concept-based selection of query expansions for spoken content retrieval,” in *Proc. 33rd int. ACM SIGIR conf. on Research and development in information retrieval*, ser. SIGIR ’10. ACM, 2010, pp. 891–892.

[87] S. Rudinac, M. Larson, and A. Hanjalic, “TUD-MIR at MediaEval 2011 genre tagging task: Query expansion from a limited number of labeled videos,” in *MediaEval*, 2011.

[88] S. Rudinac, M. Larson, and A. Hanjalic, “Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval,” *International Journal of Multimedia Information Retrieval*, vol. 1, pp. 263–280, 2012.

[89] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Inf. Process. Manage.*, vol. 24, pp. 513–523, August 1988.

[90] M. Sanderson, J. Tang, T. Arni, and P. Clough, “What else is there? search diversity examined,” in *Proceedings of the 31th European Conference on Information Retrieval*, ser. ECIR ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 562–569.

[91] A. E. Savakis, S. P. Etz, and A. C. Loui, “Evaluation of image appeal in consumer photography,” in *Proceedings SPIE Human Vision and Electronic Imaging V*, vol. 3959, Jun. 2000, pp. 111–120.

[92] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro, “How useful are your comments?: analyzing and predicting YouTube comments and comment ratings,” in *Proceedings of the 19th international conference on World Wide Web*, ser. WWW ’10. New York, NY, USA: ACM, 2010, pp. 891–900.

[93] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, “Analyzing and predicting sentiment of images on the social web,” in *Proceedings of the international conference on Multimedia*, ser. MM ’10. New York, NY, USA: ACM, 2010, pp. 715–718.

[94] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, “Summarizing visual data using bidirectional similarity,” *2012 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1–8, 2008.

[95] E. Smits and A. Hanjalic, “A system concept for socially enriched access to soccer video collections,” *IEEE Multimedia*, vol. 17, pp. 26–35, 2010.

[96] C. G. M. Snoek *et al.*, “The MediaMill TRECVID 2008 semantic video search engine,” in *Proc. TRECVID Workshop*, 2008.

[97] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 4, no. 2, pp. 215–322, 2009.

[98] C. G. M. Snoek *et al.*, "The MediaMill TRECVID 2009 semantic video search engine," in *Proc. TRECVID Workshop*, 2009.

[99] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," in *Proceedings of the 14th annual ACM international conference on Multimedia*, ser. MM '06. New York, NY, USA: ACM, 2006, pp. 707–710.

[100] A. Strehl and J. Ghosh, "Cluster ensembles – a knowledge reuse framework for combining multiple partitions," *Journal on Machine Learning Research (JMLR)*, vol. 3, pp. 583–617, December 2002.

[101] J. Tang and M. Sanderson, "Evaluation and user preference study on spatial diversity," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, vol. 5993, pp. 179–190.

[102] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Academic Press, 2008.

[103] X. Tian, Y. Lu, L. Yang, and Q. Tian, "Learning to judge image search results," in *Proc. 19th ACM int. conf. on Multimedia*, ser. MM '11. ACM, 2011, pp. 363–372.

[104] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian video search reranking," in *Proc. 16th ACM int. conf. on Multimedia*, ser. MM '08. ACM, 2008, pp. 131–140.

[105] J. Urban and J. M. Jose, "Adaptive image retrieval using a graph model for semantic feature integration," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, ser. MIR '06. New York, NY, USA: ACM, 2006, pp. 117–126.

[106] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proceedings of the 18th international conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 341–350.

[107] C. J. van Rijsbergen, *Information Retrieval.* Butterworth, 1979.

[108] N. Vasconcelos and A. Lippman, "Towards semantically meaningful feature spaces for the characterization of video content," in *Proc. Int. Conf. on Image Processing*, ser. ICIP '97. IEEE Computer Society, 1997.

[109] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, "Intelligent access to digital video: Informedia project," *IEEE Computer*, vol. 29, no. 5, pp. 46–52, 1996.

[110] C. Whissell, "Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language," *Psychological Reports*, vol. 105, pp. 509–521, October 2009.

[111] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 347–354.

[112] S. Winkler, "Visual fidelity and perceived quality: Towards comprehensive metrics," in *in Proc. SPIE Human Vision and Electronic Imaging Conference*, ser. Lecture Notes in Computer Science, vol. 4299. SPIE, 2001, pp. pp 114–125.

[113] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. S. Jones, "Effects of out of vocabulary words in spoken document retrieval," in *Proc. 23rd annual int. ACM SIGIR conf. on Research and development in information retrieval*, ser. SIGIR '00. ACM, 2000, pp. 372–374.

[114] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. 14th Int. Conf. on Machine Learning*, ser. ICML '97. Morgan Kaufmann Publishers Inc., 1997, pp. 412–420.

[115] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens, "Graph nodes clustering with the sigmoid commute-time kernel: A comparative study," *Data Knowl. Eng.*, vol. 68, pp. 338–361, March 2009.

[116] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow, "Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval," in *Proc. 28th annual int. ACM SIGIR conf. on Research and development in information retrieval*, ser. SIGIR '05. ACM, 2005, pp. 512–519.

Summary

Advancing the relevance criteria for video search and visual summarization

To facilitate finding of relevant information in ever-growing multimedia collections, a number of multimedia information retrieval solutions have been proposed over the past years. The essential element of any such solution is the relevance criterion deployed to select or rank the items from a multimedia collection to be presented to the user. Due to the inability of computational approaches to interpret multimedia items and their semantic relations in the same way as humans, the research community has mainly focused on the relevance criteria that can be handled by the modern computers, e.g., finding images or videos depicting a particular object, setting or event. However, in practice the user information needs are often specified at a higher semantic (abstraction) level, which creates a strong need for multimedia information retrieval mechanisms operating with more complex relevance criteria, such as those referring to topicality, aesthetic appeal and sentiment of multimedia items.

By considering the practical use-cases associated with different types of multimedia collections, we investigate in this thesis the possibilities of enabling video search and visual summarization based on the relevance criteria defined at a higher semantic level. To start with, we address the problem of video search at the level of semantic theme (general topic, subject matter) in the setting of an unlabeled professional video collection. For this purpose, we propose a retrieval framework based on the query performance prediction principle that makes use of the noisy output of

automatic speech recognition and visual concept detection. We demonstrate that valuable information about the semantic theme of a video can be automatically extracted from both its spoken content and the visual channel, which makes the effective retrieval within the proposed framework possible despite the presence of noise and the absence of suitable annotations.

The focus of the thesis then moves to the problem of visual summarization in information-rich social media environments. We first investigate the possibilities for improved computing of semantic similarities between images through a multimodal integration of resources ranging from image content and the associated social annotations to the information derived from the analysis of social network in which the images are contextualized. Building on the outcomes of this investigation and inspired by the prospect of using social media in tourist applications, we then propose an approach to automatic creation of visual summaries composed of community-contributed images and depicting various aspects of a selected geographic area. Although the proposed visual summarization approach is proven effective in yielding a good coverage of a targeted geographic area, like most approaches presented in related work, it suffers from a drawback that the user judgment about image suitability for the visual summary is not directly incorporated in the summarization algorithm. This observation inspires probably the most daring research question addressed in the thesis, namely, whether it is possible to learn to automatically identify images that the humans would select if asked to create a visual summary. We give a positive answer to this question and present an image selection approach that makes use of reference visual summaries obtained through crowdsourcing and a versatile image representation that goes beyond the analysis of image content and context to incorporate an analysis of their aesthetic appeal and the sentiment they evoke in the users. Finally, we address the problem of automatic evaluation of the quality of visual summaries and image sets in general, first by using the image metadata only and then based on the human-created references.

In conclusion, with this thesis we believe to have pushed the boundaries of relevance criteria that can be deployed in automated multimedia information retrieval systems by demonstrating that the video search and visual summarization can be performed at a higher semantic level. We also show, however, that the effective deployment of advanced relevance criteria requires innovative and unconventional multimedia representation

for improved capturing of semantic similarities between multimedia items. Additionally, we demonstrate that properly addressing the user information needs often requires a much more complex mix of relevance criteria than commonly assumed and prove that learning their interplay is possible. Finally, we point out that social media analysis and the emerging technologies such as e.g., crowdsourcing show a great promise in better understanding and automatically modeling the actual user information needs and the way the users interpret and interact with multimedia.

Stevan Rudinac

Samenvatting

Ontwikkeling van geavanceerde relevantiecriteria voor het doorzoeken van video en creëren van visuele samenvattingen

In de afgelopen jaren zijn verschillende technieken op het gebied van multimedia information retrieval voorgesteld die het vinden van relevante informatie in voortdurend groeiende multimediacollecties gemakkelijker moeten maken. Een essentieel onderdeel van deze technieken is het relevantiecriterium dat gehanteerd wordt om de items in een multimediacollectie te rangschikken voor presentatie aan een gebruiker. Aangezien automatische rekenmethoden niet in staat zijn om multimedia-items en hun semantische verbanden net zo te interpreteren als mensen, heeft de onderzoeksgemeenschap zich vooral beziggehouden met relevantiecriteria die door moderne computers te hanteren zijn, zoals het vinden van afbeeldingen of video's die een specifiek object of een bepaalde omgeving of gebeurtenis weergeven. In de praktijk liggen informatiebehoeften van gebruikers echter op een hoger semantisch (abstractie)niveau, wat de sterke noodzaak creëert voor technieken in multimedia information retrieval die complexere relevantiecriteria hanteren, zoals het globale onderwerp, de esthetische aantrekkelijkheid en het sentiment van multimedia-items.

Door praktische use cases te behandelen die met verschillende types multimediacollecties geassocieerd zijn, bestuderen we in deze scriptie de mogelijkheden om het doorzoeken van video en het creëren van visuele samenvattingen te baseren op relevantiecriteria die op een hoger semantisch niveau gedefinieerd zijn. Allereerst behandelen we het probleem van het

doorzoeken van video op het niveau van het semantische thema (algemeen onderwerp, inhoud) in de context van een ongeannoteerde professionele videocollectie. Hiervoor stellen we een retrievalframework voor, gebaseerd op het principe van query performance prediction, waarbij gebruik wordt gemaakt van de ruisige resultaten van automatische spraakherkenning en detectie van visuele concepten. We tonen aan dat waardevolle informatie over het semantische thema van een video automatisch uit zowel de gesproken inhoud als het visuele kanaal gehaald kan worden, wat effectieve retrieval binnen het voorgestelde framework mogelijk maakt, ondanks de aanwezigheid van ruis en de afwezigheid van geschikte annotaties.

De focus van de scriptie verplaatst dan naar het probleem van het creëren van visuele samenvattingen in informatierijke omgevingen van sociale media. Allereerst bestuderen we de mogelijkheden om het berekenen van semantische gelijkenissen tussen afbeeldingen te verbeteren via multimodale integratie van informatiebronnen, reikend van de afbeeldingsinhoud met bijbehorende sociale annotaties tot informatie afgeleid uit de analyse van het sociale netwerk waarin de afbeeldingen gecontextualiseerd zijn. Voortbouwend op de uitkomsten van dit onderzoek, en geïnspireerd door het vooruitzicht om sociale media in toeristenapplicaties te gebruiken, stellen we vervolgens een aanpak voor om op automatische wijze visuele samenvattingen te creëren, samengesteld uit afbeeldingen die door de gemeenschap zijn gemaakt, en verschillende aspecten van een geselecteerd geografisch gebied weergeven. De voorgestelde aanpak voor visuele samenvatting is effectief, in de zin dat de gewenste geografische gebieden goed afgedekt zijn. Zoals ook bij de meeste benaderingen in gerelateerd onderzoek het geval is, geldt echter het nadeel dat gebruikersoordelen over de geschiktheid van een afbeelding voor de visuele samenvatting niet direct in het samenvattingsalgoritme zijn opgenomen. Deze observatie inspireert de mogelijk meest gewaagde onderzoeksvraag van de scriptie: de vraag of het mogelijk is om automatisch afbeeldingen te leren identificeren die mensen zouden selecteren wanneer ze een visuele samenvatting zouden creëren. Wij geven een positief antwoord op deze vraag, en presenteren een techniek voor afbeeldingsselectie die gebruik maakt van visuele referentiesamenvattingen, verkregen via crowdsourcing, en een veelzijdige afbeeldingsrepresentatie die verder gaat dan de analyse van beeldinhoud en context, en een analyse meeneemt van de esthetische aantrekkelijkheid en het sentiment dat bij gebruikers opgeroepen wordt. Tot slot bespreken we het probleem van automatische evaluatie van de kwaliteit van visuele samenvattingen en beeldverzamelingen in het algemeen, allereerst uitsluitend gebruikmakend van metadata van de afbeeldingen, en vervolgens gebaseerd op door mensen gecreëerde referenties.

Samenvattend: in deze scriptie menen we de grenzen van relevantiecriteria voor multimedia-information-retrievalsystemen te hebben verlegd, door aan te tonen dat het doorzoeken van video en het creëren van visuele samenvattingen op een hoger semantisch niveau gedaan kan worden. We tonen echter ook aan dat de effectieve toepassing van geavanceerde relevantiecriteria innovatieve en onconventionele multimediarepresentaties nodig heeft om de semantische gelijkenissen tussen multimedia-items te kunnen vastleggen. Bovendien tonen we aan dat het op de juiste wijze aanpakken van informatiebehoeften van gebruikers vaak een veel complexere combinatie van relevantiecriteria nodig heeft dan gebruikelijk wordt aangenomen, en bewijzen we dat het mogelijk is om de interacties hiertussen te leren. Tot slot wijzen we erop dat analyse van sociale media en nieuw opkomende technologieen zoals bijvoorbeeld crowdsourcing veelbelovend zijn voor het beter begrijpen en automatisch modelleren van daadwerkelijke informatiebehoeften van gebruikers, en de manieren waarop gebruikers multimedia interpreteren en ermee omgaan.

Stevan Rudinac

Acknowledgements

Many people have inspired and helped me along the course of my PhD journey and, while I am deeply grateful to all of them, here, I would like to mention a few in particular.

First and foremost, I would like to express gratitude to my supervisors Alan Hanjalic, Inald Lagendijk and Martha Larson, for without them this thesis would not have been what it is, and I would not have enjoyed working on it as much as I did. I would particularly like to thank Alan because were it a problem from academic or personal sphere I needed an advice about, his doors were always open.

I am also grateful to David Tax, Marco Loog and Bob Duin from the Pattern Recognition lab, with whom I had many fruitful research discussions and whose advice helped me at various stages of my PhD.

What made my time spent at TU Delft particularly memorable were the fellow researchers from the MSP, PRB and CGV groups as well as our great support staff. I would like to thank them all for always being more than colleagues to me. I owe special thanks to Maarten Clements and Cynthia Liem for translating my propositions and thesis summary into Dutch.

Conducting research within PetaMedia Network of Excellence involved a close collaboration with a number of research institutes across Europe, including TU Berlin in Germany, EPFL and University of Geneva in Switzerland, Queen Mary, University of London in UK and Dublin City University in Ireland. Big thanks go to PetaMedia partners for many joint research activities and all the fun we had.

After joining the NFI, I still had to spend many nights and weekends working on the thesis. I am grateful to the colleagues from the NFI for their understanding and support in that period.

Finally, I owe a debt of gratitude to my family and friends, without whom I would have not come this far. I would especially like to thank my dear wife Maja for being my biggest inspiration and support through the years and for making me a better person.

Curriculum vitae

Stevan Rudinac was born on September 24, 1981 in Foča, Bosnia and Herzegovina, Yugoslavia. In 2000 he graduated as first in class from the Military Grammar School in Belgrade, Serbia. He received a degree of Graduate Engineer of Electrical Engineering from the Faculty of Electrical Engineering, University of Belgrade in 2006.

In the period 2006-2007 Stevan worked as a research assistant in the field of content-based image and video retrieval at the Innovation Center of the Faculty of Electrical Engineering in Belgrade. Subsequently, from 2007 to 2008 he conducted research on robot and machine vision at the Dynamics and Control group of the Eindhoven University of Technology.

In November 2008, he started his PhD research in the Multimedia Signal Processing group of the Delft University of Technology, under supervision of Prof. dr. Alan Hanjalic, Dr. Martha Larson and Prof. dr. ir. Inald Lagendijk. The results of his research on multimedia information retrieval with focus on video search and visual summarization are presented in this thesis.

Since November 2012, he has been employed as a forensic scientist at the Netherlands Forensic Institute.