1 | Introduction

1.1 Content-based image retrieval

The way we share information with others about the world around us has been undergoing radical changes during the past few decades. With the ubiquity of camera sensors in handheld devices and the increasing availability of mobile internet connection, images have become the primary medium for storing and conveying information, thanks to their efficiency. A photograph can be taken within seconds and still stores all visual information available to the photographer. This includes factual information such as objects with their individual states as well as the relationships and interactions among them. The factual information captured in images furthermore comprises events happening in the scene and activities performed by the objects. However, images also capture and convey information that can not be expressed by text easily, such as mood and emotions. As a reflection of the real world, a photograph preserves this raw information before it undergoes any subjective process of interpretation.

Therefore, web images are a treasure chest of information for non-scientific and scientific purposes alike. To give one example, environmental scientists have recently discovered social media images as a valuable source of information for disaster management and the analysis of flooding events (Poser and Dransch, 2010; Fohringer *et al.*, 2015; Rosser *et al.*, 2017; Barz *et al.*, 2018a). Especially in such hazardous situations, traditional sensors such as water gauges are prone to failures (Poser and Dransch, 2010). Furthermore, they are usually coarsely

distributed, such that interpolation of the data between sensors based on prior knowledge about the terrain is necessary (Apel *et al.*, 2009). These assumptions about the terrain, however, do not always hold, *e.g.*, in the case of dike breaches. Images posted on social media platforms, on the other hand, are abundant nowadays in the case of such an event and can be used to derive missing data such as the extent of the flooding (Rosser *et al.*, 2017), the grade of pollution, and the approximate inundation depth (Fohringer *et al.*, 2015). Besides improving the reliability and resolution of the sensor data, photographs can even provide additional information not available from sensors (Poser and Dransch, 2010; Fohringer *et al.*, 2015).

However, the abundance of images on the web is both a blessing and a curse: finding those images that contain the relevant or the most useful information becomes increasingly difficult and calls for sophisticated automated methods. This task of finding a certain set of images in a large database is known as *image retrieval*. Commercial engines such as *Google Image Search* or *Bing Images* have traditionally been approaching this task by relying on textual keywords that are provided by the user and describe the wanted image. These keywords are matched against the text surrounding the images on the web page on which they have been found.

While this allows reusing text-based information retrieval techniques and indices of web documents for the purpose of image retrieval, textual queries fall short in many scenarios due to the ambiguity and semantic richness of images. Figure 1.1 illustrates this phenomenon by means of an example. The image depicted there can be described from a variety of perspectives: the semantic content of the image, its artistic style, the emotions it evokes in the observer, or meta-information about the image itself. Depending on their background and the situational context, different observers will perceive and interpret this image differently. However, most images on the web are not exhaustively described in their surrounding text, for mainly two reasons: First, it is often difficult, if not impossible, to enumerate all aspects of an image explicitly, due to the potentially infinite amount of possible interpretations. Secondly, it is not necessary to do so, since most facets of an image are directly available to the viewer

OBJECTS

Maid ≺ Woman ≺ Person
Black dress
Wardrobe ≺ Furniture
Window
Liselund Castle ≺ Castle

ACTIVITIES

Daydreaming
Looking out of the window

MOOD

Melancholic
Feeling locked in

SCENE

Old-fashioned room ≺
Sunlit room ≺ Room ≺ Indoor
Woman in front of window next to wardrobe

META

„The Dream Window in the Old Liselund Castle“
≺ Painting by G. Achen ≺
≺ Oil on canvas ≺ Painting ≺ Artwork

Figure 1.1: *An example for the ambiguity and semantic richness of images. All concepts listed on the right-hand side could be used to describe the image to the left, while different observers will pay attention to different subsets of these aspects. Moreover, some concepts can be organized hierarchically, indicated by the "≺" sign, which designates the hyponomy ("is-a") relationship.*

by simply looking at it. The textual description therefore focuses most often on the meta-information that is not encoded in the image itself, such as its author. The image shown in Fig. 1.1, for example, would probably be described as a photographic reproduction of the painting "The Dream Window in the Old Liselund Castle" by Georg Achen. This would prevent this image from being found by users searching for images of a woman looking out of a window, images showing the activity "daydreaming", or images with a melancholic atmosphere.

An ideal image retrieval engine hence needs to analyze the image content itself, without relying on textual information. This also applies to the way the user formulates the query: Instead of keywords, the query itself can consist of one or more images. The users are hence relieved from the burden of translating their mental picture of what they are searching for into language, which can be highly non-trivial. Imagine, for example, the task of searching for paintings in the distinct

style of a certain artist. In this case, it is much easier to provide an example painting than a textual description of all aspects of the artist's style. An example image is, thus, a powerful device for expressing even complex search queries in compact form.

This approach of searching for images similar to an example query based on their actual content instead of a textual description is commonly referred to as *content-based image retrieval (CBIR)* (Smeulders *et al.*, 2000). The first works on this subject date back to 1992/93 (Kato *et al.*, 1992; Niblack *et al.*, 1993), when neither online image platforms nor social media were on the horizon and consumer-grade digital cameras were still far from being widespread. These early systems hence expected the user to provide a rough hand-drawn sketch of a certain image for quickly retrieving images from databases of digitized works of art. According to Smeulders *et al.* (2000), it took until 1997, though, before research on CBIR got up to speed. A second wave of novel techniques was initiated by the work of Sivic and Zisserman (2003), who applied established methods for text retrieval to images. In 2014, another paradigm shift towards re-using image features learned end-to-end using deep learning techniques led to the most recent incarnation of CBIR systems (Babenko *et al.*, 2014; Razavian *et al.*, 2014). Today, CBIR is successfully employed for a variety of applications, including

- similarity-based photo search for the web (Hu *et al.*, 2018b), image platforms (Clayton *et al.*, 2017; Zhai *et al.*, 2019), and digital libraries (Zhu *et al.*, 2000),
- visual product search also known as "shop the look" (Yang *et al.*, 2017; Hu *et al.*, 2018b; Zhai *et al.*, 2019),
- medical applications, where efficient retrieval is crucial for analyzing the large amount of data generated by medical imaging devices (Qayyum *et al.*, 2017),
- biodiversity research by searching for images showing occurrences of certain species (Sheikh *et al.*, 2011; Freytag *et al.*, 2015),
- discovery of handwritten documents in the handwriting of a certain author (Christlein *et al.*, 2019),
- classification in an open world or with limited training data (Sung *et al.*, 2018; Freytag *et al.*, 2015; Göring *et al.*, 2014).

1.2 Instance vs. category retrieval

Assessing the similarity of two images is the core task of CBIR. Similarity, however, is a subjective and vague concept. One user of a CBIR system may consider a pair of images to be similar which another user does not, depending on the objective they pursue with their search. For example, the first user could be interested in finding copies of a certain photo of a person on the web, which only vary with respect to their resolution, contrast, the section of the original photo being shown *etc*. The second user could be searching for different images of the same person and a third user might be interested in finding all images that are portraits of any person. Similarity is, thus, a graded concept ranging between the two poles of a *narrow* and a *broad image domain* (Smeulders *et al.*, 2000). Depending on where the search objective lies on this continuum, three major types of CBIR tasks can be distinguished (examples are given in Fig. 1.2):

Duplicate retrieval searches for images with the same semantic content. These variants originated from the same photo but might have been post-processed differently with regard to cropping, scaling, brightness, contrast *etc*.

Instance retrieval searches for images that contain the same instance of an object, *i.e.*, a person or a building. Thanks to its nature as a well-defined but non-trivial task with a clear ground-truth, this is the most extensively studied CBIR sub-task (noteworthy examples include Sivic and Zisserman, 2003; Jégou *et al.*, 2010; Jégou and Zisserman, 2014; Husain and Bober, 2017; Babenko *et al.*, 2014; Babenko and Lempitsky, 2015; Tolias *et al.*, 2016). A handful of established datasets is available for this task (most notably Jégou *et al.*, 2008; Philbin *et al.*, 2007, 2008; Radenović *et al.*, 2018) and significant progress has been made during the past few years (Gordo *et al.*, 2017; Revaud *et al.*, 2019). This problem can hence be considered solved to a large extent.

Category retrieval covers the remaining spectrum broader than instance retrieval and aims for finding images belonging to the same category as the query. It is important to note that the set of possible

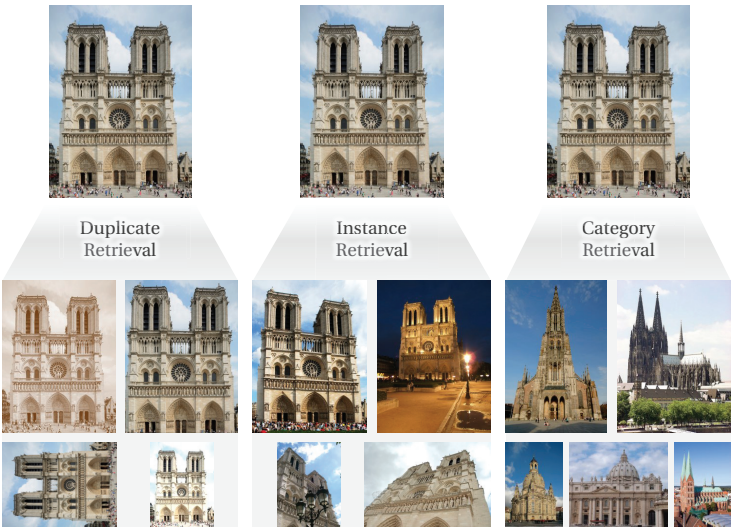

Figure 1.2: *Examples for three different sets of images to be retrieved given the same query depending on the CBIR task.*

categories is limited by nothing but the imagination of the user and that a single image usually belongs to a surprisingly high number of categories at once (see Fig. 1.1). Thus, the exact search objective of the user cannot be determined based on the query image alone and will almost certainly also vary between users, even for the same query. Therefore, approaches to this problem often comprise interaction with the user to adapt the similarity measure used by the system to that in the user's mind (*e.g.*, Cox *et al.*, 2000; Deselaers *et al.*, 2008; Barz and Denzler, 2018a; Mehra *et al.*, 2018).
This CBIR type has only been approached sporadically in recent literature (*e.g.*, Yu *et al.*, 2017; Piras and Giacinto, 2017; Hu *et al.*, 2018b), but is relevant for a variety of applications (some examples have been mentioned in Section 1.1). Simply adopting established techniques from the more well-studied field of instance retrieval often falls short, as illustrated in Fig. 1.3: All methods tuned towards instance retrieval perform worse for the more general task of category retrieval than the

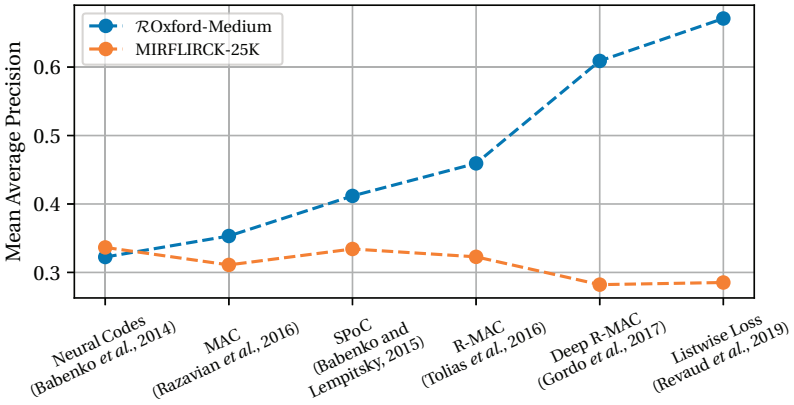

Figure 1.3: *Evolution of the state of the art in CNN-based instance retrieval, evaluated on an instance retrieval (Revisited Oxford5k, Philbin* et al.*, 2007; Radenović* et al.*, 2018) and a category retrieval dataset (MIRFLICKR-25K, Huiskes and Lew, 2008). Details about this experiment can be found in Appendix A.5.*

least complex, but most generic solution from 2014. On the other hand, this facet of image retrieval probably arouses the greatest public interest due to its wide applicability. This is reflected, for instance, by the press and media coverage regarding some of the work presented in this thesis (*e.g.*, Deutschlandfunk, 2019). More research on methods optimized for category retrieval is hence highly desirable and, therefore, the focus of this thesis.

1.3 Challenges

Content-based image retrieval, and category retrieval in particular, pose a plethora of challenges, which have been attracting researchers from a variety of backgrounds. We list some of the most relevant challenges in the following, but this list is certainly not exhaustive.

- At first glance, one might easily be tricked into thinking that category retrieval can be solved using methods and models from the research

area of *image classification* by classifying both query and database images and matching their predicted labels. However, image classification techniques operate under the assumption that each image can be deterministically assigned to a subset of a finite set of labels that is known in advance. Frequently, this label subset is even limited to a single label per image. In the context of category retrieval, on the other hand, we are not only faced with **multiple labels for a single image** being the typical scenario, but the **set of possible labels is also theoretically infinite** and not known in advance. This requires a much more **generic and expressive feature representation** for comparing images than classification.

- **There is no ground-truth.** As opposed to instance retrieval, where it is usually indisputable whether two images show the same instance of, *e.g.*, a building, the relevance of an image with respect to a certain query varies between users in the case of category retrieval. This **complicates the quantitative evaluation** and comparison of category retrieval methods.

- The lack of ground-truth annotations implies the **absence of a dedicated training phase**. Category retrieval methods hence need to be trained on proxy tasks such as classification (*e.g.*, Babenko *et al.*, 2014; Hu *et al.*, 2018b; Zhai and Wu, 2019) or in an unsupervised way (*e.g.*, Radenović *et al.*, 2018; Noroozi and Favaro, 2016).

- **Relevance is not a binary but a graded phenomenon** (Smeulders *et al.*, 2000), calling for more sophisticated evaluation metrics and learning techniques from the research field of *learning to rank* (*e.g.*, Gordo *et al.*, 2017; Revaud *et al.*, 2019).

- The **spatial layout** of the scene and the **relationships between objects** may be important for some search objectives (see Section 2.5.5).

- A nearest-neighbor search in **huge datasets** with complex features needs to be performed within **user-acceptable time**.

- Most semantic concepts are **hierarchically organized** (see Fig. 1.1) and the level of abstraction that the user is aiming for is unclear. For

example, given a query image of a poodle, the user could be searching for images of poodles only or images of dogs or even animals in general. It might hence be desirable to organize the retrieval results by their degree of **semantic similarity to the query according to a taxonomy**. We propose such a method in Chapter 4 and show that it improves the semantic consistency of CBIR results considerably.

- Due to the aforementioned **ambiguity**, mechanisms for the **interactive manipulation** of the results are mandatory to acquire sufficient context (Smeulders *et al.*, 2000). This can even require dynamic image features and similarity measures for **adapting the retrieval system** to the user's needs on the fly. We propose three novel approaches to this problem for different types of interaction in Chapter 6.

1.4 Interactive image retrieval

As mentioned above, interactive mechanisms are not optional but necessary for category retrieval. In contrast to duplicate or instance retrieval, the search objective cannot be determined based solely on a single query image provided by the user at the beginning. A generic CBIR engine hence needs to keep the users in the loop and involve them at several stages of the retrieval process to acquire feedback regarding the current results and hints towards the right direction in the search space. This feedback can take a variety of modalities.

The oldest but still highly effective and prominent form of user feedback is *relevance feedback*, which had already been used in the field of text retrieval for a long time (Rocchio, 1971) before it was adopted for CBIR applications in the late-1990s (*e.g.*, Picard *et al.*, 1996; Cox *et al.*, 2000). Under this regime, the system first performs an initial *baseline retrieval* based on the query image provided by the user and presents the top-scoring results. The user may then provide relevance ratings for a handful of the retrieved images, for example by flagging them as relevant or irrelevant (see Fig. 1.4a). The system then incorporates this feedback to adapt to the user's needs and presents a refined list of results. This process can be iterated multiple times until the user is satisfied with the results. Following Zhou and Huang (2003), relevance

feedback approaches can be categorized along several axes depending on the typical user needs, the application scenario, and the technical requirements of the method:

- Is the **target of the search** a certain individual image or a class of similar images? The former assumption is prevalent in some application areas such as product image retrieval (*e.g.*, Plummer *et al.*, 2019), while the latter is more typical for category retrieval.
- The **type of the user feedback** can be a set of positive images only, binary judgments as either relevant or irrelevant, or even a degree of (ir)relevance on an ordinal (Rui *et al.*, 1998) or even continuous (Kim and Chung, 2003) spectrum. The latter approach takes account of the fact that relevance is not a binary but a graded phenomenon, but very few works pursue this path. This is probably due to the impracticality of assessing the relevance of an image relative to all other images without knowing the entire dataset in advance.
- **The user** can be greedy or cooperative. A *greedy user* expects to see the best results after each feedback round and can terminate the feedback loop at any time. A *cooperative user*, on the other hand, would be willing to go through several rounds of feedback with less relevant images in order to get better results in the end.
- Since users are, in general, very different and pursue a variety of search objectives, most retrieval systems perform *short-term* or *intra-query learning*, where no information is shared between different sessions. That means, each user starts from scratch with the same baseline retrieval system. In certain scenarios, however, where users can be assumed to pursue similar search objectives, *long-term* or *inter-query learning* can be beneficial, where the system adapts and improves continuously using all feedback from past sessions.

Some important existing relevance feedback techniques are explained in detail in Section 2.7. They all fall into the category of short-term learning with binary relevance feedback and a greedy user. The greediness assumption, however, prevents relevance feedback methods from exploiting their full potential: Since the user is presented with the top-scoring images at each round, which the system already considers as